$$\delta \int_{t_1}^{t_2} L(q, \dot{q}, t)\, dt = 0$$

VARIATIONAL METHODS IN MATHEMATICAL PHYSICS

S. G. MIKHLIN

$$J[u] = \int_{\Omega} L(x, u, \nabla u)\, d\Omega \to \min$$

INTERNATIONAL SERIES OF MONOGRAPHS IN
PURE AND APPLIED MATHEMATICS
GENERAL EDITORS: I. N. SNEDDON, M. STARK AND S. ULAM

VOLUME 50

VARIATIONAL METHODS IN MATHEMATICAL PHYSICS

VARIATIONAL METHODS IN MATHEMATICAL PHYSICS

by

S. G. MIKHLIN

Translated by

T. BODDINGTON

Editorial Introduction by
Translation Editor

L. I. G. CHAMBERS
Lecturer in Applied Mathematics
University College of North Wales
Bangor

A Pergamon Press Book

THE MACMILLAN COMPANY
NEW YORK
1964

THE MACMILLAN COMPANY
60 Fifth Avenue
New York 11, N.Y.

This book is distributed by
THE MACMILLAN COMPANY
pursuant to a special arrangement with
PERGAMON PRESS LIMITED
Oxford, England

Copyright © 1964
PERGAMON PRESS LTD.

This is a translation from the original Russian *Вариатсионные методы в математической физике*, (*Variatsionnyye metody v matematicheskoi fizike*), published in 1957 by Gostekhizdat, Moscow

Library of Congress Catalog Card Number 63 – 10021

CONTENTS

Editor's Preface ix
Foreword xi
Introduction; Historical Review xv

I. THE OPERATORS OF MATHEMATICAL PHYSICS

§ 1. The formulation of the basic problems 1
§ 2. Some ancillary concepts and formulae 6
§ 3. The scalar product of functions 12
§ 4. The concepts of operator and of functional 20
§ 5. Symmetric operators 25
§ 6. Positive definite and positive-bounded-below operators 31

II. ENERGY CONVERGENCE

§ 7. The estimation of the error of an approximation and the nature of its convergence. Convergence in the mean 40
§ 8. Energy convergence 51
§ 9. The linear independence of functions 56
§ 10. Orthogonality and orthogonal series 59

III. THE ENERGY METHOD

§ 11. The minimum functional theorem 74
§ 12. Representation of a solution in the form of an orthogonal series 80
§ 13. The minimizing sequence and its convergence 83
§ 14. The Ritz method 85
§ 15. Other methods of constructing the minimizing sequence 95
§ 16. Functions with finite energy 100
§ 17. The applications of functions with finite energy. Natural boundary conditions 109
§ 18. Non-homogeneous boundary conditions 116
§ 19. The existence of a solution of the variational problem 121

IV. MAJOR APPLICATIONS OF THE ENERGY METHOD

§ 20. Boundary value problems for an ordinary differential equation 126
§ 21. Bending of a beam of variable cross-section, resting on an elastic foundation 136

§ 22. Fundamental boundary value problems for the equations of Poisson and Laplace 138
§ 23. Problems of torsion of a bar and of bending of a beam by a lateral force 151
§ 24. Equations with variable coefficients 160
§ 25. Degenerate elliptic equations; the Chaplygin equation 167
§ 26. The principle of minimum of potential energy in elasticity theory 175
§ 27. The bending of thin plates 182
§ 28. The bending of thin plates to which both normal loads and loads acting in the central plane are applied 200

V. THE EIGENVALUE PROBLEM

§ 29. The eigenvalue problem; its connexion with characteristic vibrations and stability of a system 207
§ 30. Eigenvalues and eigenfunctions of a symmetric operator 213
§ 31. Energy theorems in the eigenvalue problem 218
§ 32. Ritz's method for the eigenvalue problem 226
§ 33. Another form of Ritz's method; the case of natural boundary conditions 233
§ 34. Equations of the form $Au-\lambda Bu=0$ 236
§ 35. The eigenvalues of an ordinary differential equation 239
§ 36. The stability of a compressed rod 250
§ 37. Eigenvalues of elliptic operators 253
§ 38. The stability of a compressed plate 262
§ 39. The characteristic vibrations of an elastic body 265
§ 40. The minimax principle and its implications 270

VI. GENERALIZATION OF THE PRECEDING RESULTS

§ 41. The concept of Lebesgue integral 278
§ 42. Functional Hilbert spaces 287
§ 43. The limit process in Hilbert spaces 295
§ 44. Generalization of the concept of orthogonality 300
§ 45. General definitions of functional and operator 308
§ 46. The general formulation of the variational problem and its solution 318
§ 47. The method of minimum surface integrals 328

VII. ESTIMATES OF THE ERROR OF AN APPROXIMATE SOLUTION

§ 48. General remarks 332
§ 49. Subspaces and projections 335
§ 50. The method of orthogonal projections in Dirichlet's problem 339
§ 51. General formulation of the method of orthogonal projections 345

§ 52. Some additional considerations 350
§ 53. The Neumann problem 352
§ 54. Castigliano's principle and two-sided estimates in elasticity theory 355
§ 55. Trefftz's method 359
§ 56. The biharmonic equation. The method of anharmonic residue 365
§ 57. Generalization of Trefftz's method 369
§ 58. Application to Poisson's equation 371
§ 59. Generalization of Trefftz's method to the problem of a freely supported plate 375
§ 60. The method of M. G. Slobodyansky 380
§ 61. Upper and lower bounds for functionals 383
§ 62. Upper and lower bounds of eigenvalues 385
§ 63. Estimate of the error due to a perturbation of the equation 391

VIII. NUMERICAL EXAMPLES

§ 64. On coordinate functions 395
§ 65. The torsion of a rod of rectangular cross-section 402
§ 66. Bending of a rectangular plate, rigidly fixed at the rim 413
§ 67. Bending of semicircular plate elastically supported at the rim 419
§ 68. Evaluation of the eigenvalues of an ordinary second-order differential equation 423
§ 69. Eigenvibrations of a beam of variable cross-section 427
§ 70. Radial eigenvibrations of an elastic cylinder 435
§ 71. Plane vibrations of an elastic rectangular plate 440
§ 72. Stability of a compressed elliptic plate 445

IX. THE BUBNOV—GALERKIN METHOD

§ 73. Foundations of the method 448
§ 74. Proof of convergence for Fredholm integral equations 450
§ 75. Proof of convergence for ordinary second-order differential equations 455
§ 76. Completely continuous operators 459
§ 77. Equations involving a completely continuous operator 462
§ 78. A sufficient criterion for the convergence of the Bubnov—Galerkin method 469
§ 79. Application to ordinary differential equations 477
§ 80. The Dirichlet problem for a second-order elliptic equation 480
§ 81. The Neumann problem and the mixed problem for a second-order elliptic equation 485
§ 82. Modification of the Bubnov—Galerkin method for the case of natural boundary conditions 489

X. THE METHOD OF LEAST SQUARES

§ 83. Basis of the method 491
§ 84. Application to integral equations 499
§ 85. Application to boundary value problems with homogeneous boundary conditions 503
§ 86. Auxiliary theorems of the theory of analytic functions 504
§ 87. The Dirichlet and Neumann problems 509
§ 88. The Dirichlet problem for an ellipse 514
§ 89. The case of a piecewise smooth contour. The Dirichlet problem 515
§ 90. The mixed problem of potential theory 518
§ 91. The plane problem of elasticity theory 526
§ 92. The periodic problem of elasticity theory 530
§ 93. Stress in an elastic space bounded by a sine curve 538

XI. FINITE DIFFERENCE METHODS

§ 94. The method of networks 544
§ 95. Foundations of the method of straight lines 549
§ 96. Differential equations of the method of straight lines for the equations of Poisson and Laplace 551
§ 97. The case of a trapezoidal domain 554
§ 98. Differential equations of the method of straight lines for the biharmonic equation 563

Bibliography 567
Index 577

List of Volumes in the Series in Pure and Applied Mathematics 583

EDITOR'S PREFACE

IN THIS translation certain errors in the original have been corrected and footnotes have been incorporated into the text. The notation has also been revised in some places to conform with British usage. In the bibliography those works which in the original were referred to in Russian translations have, wherever possible, been referred to in their original form. Correspondingly, whenever an English translation is available, this is referred to as well as to the original.

I have to thank Mr. T. Boddington for a translation which needed very little editing and Mr. D. M. Platt, translation editor of the Pergamon Institute, for his kindness and help during the course of the editing.

L. I. G. CHAMBERS

FOREWORD

THE PRESENT book is the result of a considerable revision of the author's *Direct Methods in Mathematical Physics* which appeared in 1950. The revision was carried out principally with two aspects in view. First of all, the most important of the variational methods – the energy method – has been described for the fundamental problems of mathematical physics on a more elementary mathematical basis, without introducing the theory of operators in a Hilbert space. A more general point of view arising naturally from the material already collected is given only in Chapter VI. The author hopes that this rearrangement in the presentation will make the book accessible to a wider circle of readers.

Another change which seems to be very important to the author is the following. In the last few years several papers have appeared, here as well as abroad, which discuss the error of an approximate solution. The existence of such papers enables the problem of estimation of error of an approximate solution given by the energy method to be dealt with in sufficient detail for use in practical problems. A particularly long chapter has been devoted in the present book to this problem, which was only touched upon in the book *Direct Methods*.

In addition to the radical changes mentioned above, a number of important, though not so significant, changes have been made:

1. A chapter on the formulation of the fundamental problems of mathematical physics and the most important properties of the operators occurring most frequently in mathematical physics has been introduced.

2. The problem of natural boundary conditions has been considered more fully.
3. The theory of eigenvalues has been considered in considerably greater detail, a special chapter being devoted to it.
4. The number of numerical examples has been reduced. On the other hand, in most of the remaining examples the calculations include the estimation of errors.
5. The validity of the Bubnov–Galerkin method has been proved without using the theory of infinite systems.
6. The main material of Chapters II (Variational principles in mathematical physics) and III (Methods of solving variational problems) of the book *Direct Methods* is brought together in Chapter IV of the present book.
7. Certain sections of the book *Direct Methods* have been excluded to avoid a considerable increase in volume.

The author decided to change the name of the book, since almost the whole of it is devoted just to variational methods – the energy method, the method of least squares, the method of orthogonal projections, Trefftz's method, and also the Bubnov–Galerkin method which is closely connected with the energy method. Very little space has been devoted to direct methods which are not variational, i.e. methods of finite differences. As regards the method of networks, which has been adequately dealt with in monographs, the author has confined himself to an enumeration of the main results and references to the literature: as regards "the method of straight lines", it is only applied to Laplace's equation and the biharmonic equation. As far as is known to the author, the method of straight lines has been dealt with in this form only in articles in journals. For this reason, he has retained the corresponding chapter, although it stands apart from the remaining chapters.

For understanding the basic material of the present work (Chapters I—V) an acquaintance with the first two volumes of V. I. Smirnov's *Course of Higher Mathematics*, with the

exception of Chapter V, Vol. II ("Principles of Differential Geometry") is required on the part of the reader. A knowledge of some facts of the theory of functions of a complex variable and the theory of integral equations is required in certain places. In such cases references to the necessary literature will be made in the text. To simplify very frequent references we shall denote vol. I and vol. II of V. I. Smirnov's *Course* (1956 edition) by 1 and 2.

The author owes a special debt of gratitude to K. Ye. Chernin, who brilliantly carried out a number of new calculations for the present book, and to G. P. Akilov, who read through the entire book in manuscript form and made many valuable comments.

S. G. MIKHLIN

INTRODUCTION: HISTORICAL REVIEW

THE PROBLEMS of mathematical physics and also of elasticity theory, hydrodynamics etc. usually lead to partial differential equations, more rarely to ordinary differential equations, which have to be integrated under the initial or boundary conditions corresponding to the given problem. From the practical viewpoint the evaluation of numerical, albeit approximate, values of the required quantities is of considerable interest. Perhaps the best devices for this purpose are the so called *direct methods*. It is difficult to give a precise definition of this group of methods. According to S. L. Sobolev* *direct methods are those methods for the approximate solution of the problems of the theory of differential and integral equations which reduce these problems to finite systems of algebraic equations.* Of a number of definitions with which the author is familiar this is the most common and the most complete; we shall retain it, although some of the methods with which we shall be concerned below and which we shall consider to be direct are difficult to include in the above definition (for example, Kantorovich's method (cf. § 15)).

In theory and in the practical applications of direct methods we often come across a fact which is of prime importance. This is that in many cases it is possible to replace the problem of integrating a differential equation by the equivalent problem of seeking a function which gives a minimum value to some integral. Problems of such a type are called *variational problems;* thus the fact referred to above implies that in many cases the problem of integrating a differential equation can

* cf. The course by S. L. Sobolev [1], 1st edition, 1947.

be replaced by some equivalent variational problem. Thus, for example, under the usual boundary conditions it is possible to reduce the integration of the equations of static elasticity theory to the finding of a minimum of the potential energy of the elastic body. The methods which allow us to reduce the problem of integrating a differential equation to the equivalent variational problem are usually called *variational methods*. The most important of the variational methods is known in the technical literature as the *energy method*. This designation seems fortunate to us and we shall adhere to it. The energy method is discussed in more detail in Ch. III—VI.

It happens that many direct methods are conveniently applied not immediately to the differential equation but to the variational problem equivalent to it. This applies above all to the well known Ritz method which is an approximate method for the solution of variational problems. We have thus explained the close connexion which exists between direct and variational methods.

Historically the first variational method was formulated in the form of the so called "Dirichlet principle", according to which of all the functions assuming specified values on the boundary of some domain Ω the function which gives the so called "Dirichlet integral"

$$\int_{\Omega}\int \left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]\mathrm{d}x\mathrm{d}y \tag{1}$$

the least value is harmonic in Ω (here for simplicity we have considered the plane problem only).

The Dirichlet principle was widely used by Riemann in his work on the theory of functions of a complex variable. Riemann, evidently, considered that a function giving the Dirichlet integral a minimum value exists. This evoked a critical observation from Weierstrass [1] who showed by a simple example that *a minimum value of the integral need not occur*.

Thus Weierstrass considered the following problem (in fact we have simplified it somewhat by stripping it of inessential details): of the functions which are continuous and continuously differentiable on the segment $-1 \leq x \leq 1$ and which satisfy the conditions

$$y(-1) = -1, \quad y(+1) = +1, \tag{2}$$

at the ends of this segment, we are required to find the function which gives a minimum value to the integral

$$J(y) = \int_{-1}^{+1} x^2 y'^2 dx. \tag{3}$$

The value of the integral $J(y)$ depends upon the choice of the function $y(x)$; obviously, $J(y) \geq 0$ since integral (3) is bounded below (cf. 1, 39) and thus has an exact lower bound (cf. 1, 42). We shall prove that this bound equals zero; for this it suffices to satisfy ourselves that it is possible to find a function $y(x)$ which satisfies the conditions given above and such that it gives a value less than any prespecified positive number to the integral $J(y)$. We specify a number $\varepsilon > 0$ and put

$$y_\varepsilon = \frac{\arctan \frac{x}{\varepsilon}}{\arctan \frac{1}{\varepsilon}}.$$

Obviously the function y_ε is continuous and continuously differentiable in the interval $-1 \leq x \leq 1$ and satisfies conditions (2). Moreover,

$$J(y_\varepsilon) = \int_{-1}^{1} x^2 y_\varepsilon'^2 dx < \int_{-1}^{1} (x^2 + \varepsilon^2) y_\varepsilon'^2 dx =$$

$$= \frac{\varepsilon^2}{\left(\arctan \frac{1}{\varepsilon}\right)^2} \int_{-1}^{1} \frac{dx}{x^2 + \varepsilon^2} = \frac{2\varepsilon}{\arctan \frac{1}{\varepsilon}}.$$

Hence it is clear that $J(y_\varepsilon)$ can be as small as required if ε is sufficiently small. By definition the exact lower bound of the integral $J(y)$ equals zero. However, no function exists which is continuous and continuously differentiable in the interval $-1 \leq x \leq 1$ and satisfies conditions (2) which also causes the integral $J(y)$ to vanish. In fact, let $J(y) = 0$. Since the integrand is non-negative it must be identically equal to zero. Then $y' = 0$ and $y = \text{const}$ which contradicts conditions (3).

Thus it can happen that the integral does not reach its lower bound and the corresponding variational problem then has no solution. This circumstance put the Dirichlet principle in doubt.

Considerably later Hadamard gave an example of another kind which showed that the Dirichlet principle may not be valid for the simple reason that a harmonic function can make the Dirichlet integral infinite. We now give this example.

It is readily seen that the series

$$u(x, y) = \sum_{n=1}^{\infty} \frac{\varrho^{2^{2n}}}{2^n} \cos(2^{2n}\vartheta); \quad x = \varrho \cos\vartheta, \quad y = \varrho \sin\vartheta \qquad (4)$$

converges uniformly in the circle $\varrho \leq 1$; its sum is harmonic within this circle and continuous as far as the boundary. However the Dirichlet integral of this function for a circle $\varrho \leq r < 1$ equals

$$\pi \sum_{n=1}^{\infty} r^{2^{2n}+1} \underset{r \to 1}{\to} \infty.$$

Thus it is impossible to obtain the harmonic function (4) as the solution of a variational problem on the basis of Dirichlet's principle.

The result of Weierstras's objection was that Dirichlet's principle was neglected for a long time. Interest in this principle and in the variational method in general was again aroused at the start of the twentieth century in connexion with the work of Hilbert and especially after W. Ritz [1] had published (1908) his method for the approximate solution of minimum problems; interest was particularly shown by workers in the field of the applied sciences for whom Ritz has provided a convenient tool for the solution of problems which were previously quite intractable. Many papers emerged in which approximate solutions of various problems of mathematical physics were constructed with the aid of the Ritz method. In those cases where it was possible to compare such an approximate solution with the exact one or with experimental results the coincidence was usually satisfactory.

Especially good results were obtained when a boundary value problem for an ordinary differential equation was solved.

Ritz expounded his method in the following way. Suppose the variational problem for the integral

$$J(w)=\int_a f(x,\ w,\ w',\ w'',\ \ldots,\ w^{(k)})\mathrm{d}x \tag{5}$$

is posed.

As is known this problem consists of the following: a certain class of functions which are usually known as admissible functions is considered; we must find the function in this class which gives integral (5) a lower value than any other admissible function.

We specify a certain sequence of functions

$$\psi_0,\ \psi_1,\ \psi_2,\ \ldots,\ \psi_n,\ \ldots, \tag{6}$$

which we subject to two stipulations: (1) for any values of the numerical coefficients $a_1,\ a_2\ ,\ldots,\ a_n$ the function

$$w_n=\psi_0+a_1\psi_1+a_2\psi_2+\ \ldots\ +a_n\psi_n \tag{7}$$

belongs to the class of admissibles;
(2) for no matter what admissible function w it is possible to find an integer n and the numerical coefficients $a_1,\ a_2\ ,\ldots,$ a_n such that the function w_n defined by formula (7) differs sufficiently little from the function w.

We note that stipulation (2) usually called the *completeness condition* is needed in a more precise formulation. Ritz makes the necessary improvement in the precision for the concrete problems (see below) to which he applied his method. Ritz called the functions (6) *coordinate functions.*

In integral (5) we substitute w_n for w, then J is transformed into a function of the numbers $a_1,\ a_2\ ,\ldots,\ a_n$. We shall select these numbers so that $J(w_n)$ assumes a minimum value; as is well known, it is necessary for this that $a_1,\ a_2,\ \ldots,\ a_n$ satisfy the conditions

$$\frac{\partial J(w_n)}{\partial a_1}=0,\quad \frac{\partial J(w_n)}{\partial a_2}=0,\quad \ldots,\quad \frac{\partial J(w_n)}{\partial a_n}=0. \tag{8}$$

By solving equations (8) and substituting the obtained values $a_1, a_2, \ldots, a_n$ into (7) we obtain the function w_n which Ritz considers as the approximate solution of the variational problem.

The fact that w depends only upon the one variable x is obviously of no importance: the computational process is the same when w is a function of any number of independent variables and correspondingly $J(w)$ is a multiple integral.

In the general case the construction of the approximate solution w_n meets with fairly considerable computational difficulty, i.e. with the necessity of solving system (8). However, if the integrand f in (5) is a polynomial of the second degree in w and its derivatives then equations (8) are linear and their solution is a matter of comparative simplicity. This happens whenever the variational problem is connected with some *linear* problem of mathematical physics.

In connexion with the Ritz method it is necessary to answer the following question: to what extent is it possible to consider w_n as an approximation to the true solution of the variational problem? This question of the substantiation of the method is scarcely soluble in general form. Below, in Ch. III – VI a substantiation of the Ritz method will be given for a fairly wide class of the linear problems of mathematical physics and of elasticity theory.

Ritz himself well understood the necessity for substantiating his method. In [1] Ritz briefly formulates his method in general form and then investigates its convergence in detail for three problems: (1) the deflexion of an elastic plate which is rigidly fixed at the edge and under the action of a normal pressure; (2) Dirichlet's problem for Laplace's equation; (3) the boundary value problem for an ordinary, second-order differential equation. At the end of the paper he gives an application of the method to the problem of the eigenvibrations of a string. Here the author does not dwell on the question of the validity of applying his method, but shows that the approximate values of the eigenfrequencies of the

string obtained by this method are extremely close to the well known exact values. We note here that in application to problems of the theory of vibrations the Ritz method is a far-reaching generalization of the so called "Rayleigh Method" (cf. Rayleigh [1]). The computationally more complex problem of the eigenvibrations of a square plate was considered by Ritz in his paper [2].

We shall set out at greater length one of the problems studied by Ritz, i.e. the problem of the plate. At various places we shall make unimportant deviations from Ritz's reasoning or notation. Moreover, whenever it is necessary to be more precise, various additions are made. We shall also omit some further not very obvious details of the proof to enable us to distinguish the fundamental ideas clearly.

Let the plate occupy the domain Ω of the (x, y) plane in the equilibrium state and let S be the boundary of this domain. Let the plate be rigidly fixed at its edge and subject to the influence of a normal pressure of intensity $q(x, y)$. The normal deflexion of the plate which we shall denote by $w(x, y)$ satisfies the well-known differential equation due to Sophie Germain

$$\nabla^4 w = f(x, y); \quad f(x, y) = \frac{1}{D} q(x, y) \tag{9}$$

within the domain Ω and the boundary conditions

$$w = 0, \quad \frac{\partial w}{\partial n} = 0 \tag{10}$$

on the boundary S. Here n is the direction of the normal to the boundary and D is the rigidity of the plate towards deflexion. It is known (for more detail see below, § 27) that the problem of integrating equation (9) under boundary conditions (10) can be reduced to the following variational problem: from the functions satisfying boundary conditions (10) (these are the admissible functions for our variational problem) we are required to find that which gives the least value to the integral

$$J(w)=\int_{\Omega}\int\left[\frac{1}{2}(\nabla^2 w)^2-fw\right]\mathrm{d}\Omega. \tag{11}$$

Integral (11) differs only by a constant factor from the potential energy of the deformed plate. In solving the variational problem just posed we assume that $f(x, y)$ is continuous and continuously differentiable within and on the boundary of the domain of the plate.

Let (x, y) and (ξ, η) be two points in this domain and let r be the distance between them.

We put $w=w_1+w_2$ in (11), where

$$w_1=-\frac{1}{8\pi}\int_{\Omega}\int r^2 \ln r f(\xi, \eta)\mathrm{d}\xi\mathrm{d}\eta. \tag{12}$$

By means of simple identity transformations it is established that for any admissible function $w(x, y)$ there occurs the identity

$$J(w)=J_0+\frac{1}{2}\int_{\Omega}\int(\nabla^2 w_2)^2\mathrm{d}x\mathrm{d}y, \tag{13}$$

where J_0 is some constant. Since the integral in (13) is obviously non-negative,

$$J(w)\geq J_0. \tag{14}$$

Thus the integral $J(w)$ is bounded below if the function $w(x, y)$ belongs to the class of admissibles, and thus this integral has an exact lower bound. Our variational problem consists of seeking an admissible function which gives the integral $J(w)$ a value equal to its exact lower bound.

In order to construct an approximate solution of the problem we select a sequence of coordinate functions

$$\psi_1(x, y),\ \psi_2(x, y),\ \ldots,\ \psi_n(x, y),\ \ldots, \tag{15}$$

subject to the following stipulations:

(1) the functions $\psi_n\ (x, y)$ and their derivatives* of the form

* Ritz calls these derivatives *principal derivatives.*

$$\frac{\partial^{k+l}\psi_n}{\partial x^k \partial y^l}, \quad k \leq 3, \ l \leq 3$$

are continuous within and on the boundary of the domain Ω;
(2) the coordinate functions satisfy the boundary conditions (10);
(3) let ζ be an arbitrary function which is continuous together with its principal derivatives within and on the boundary of the domain Ω and is identically equal to zero outside some rectangle ϱ which with its boundary lies entirely within Ω. Then it is possible to select an integer m and the constants $\alpha_1, \alpha_2, \ldots, \alpha_m$ such that the inequalities

$$\left|\zeta - \sum_{i=1}^{m} \alpha_i \psi_i\right| < \varepsilon, \quad \left|\frac{\partial^{k+l}\zeta}{\partial x^k \partial y^l} - \sum_{i=1}^{m} \alpha_i \frac{\partial^{k+l}\psi_i}{\partial x^k \partial y^l}\right| < \varepsilon, \ k \leq 3, \ l \leq 3, \quad (16)$$

hold, where ε is a previously specified positive number which is as small as required (*the completeness condition*).

We shall seek the approximate solution of the problem in the form

$$w_n = a_1\psi_1 + a_2\psi_2 + \ldots + a_n\psi_n,$$

where the number of terms n is chosen arbitrarily and the constants $a_1, a_2, \ldots, a_n$ are found from the condition that the quantity

$$J_n = \int_{\Omega}\int \left[\frac{1}{2}(\nabla^2 w_n)^2 - f w_n\right] dx dy$$

have a minimum value. Equations (8) in the case considered have the form

$$\sum_{k=1}^{n} A_{ik} a_k = B_i, \quad i = 1, 2, \ldots, n, \quad (17)$$

where

$$A_{ik} = \int_{\Omega}\int \nabla^2 \psi_i \nabla^2 \psi_k dx dy, \quad B_i = \int_{\Omega}\int f \psi_i dx dy. \quad (18)$$

Using the theory of quadratic forms, Ritz proves that system

(17) has one and only one solution. By substituting this solution in the expression for w_n we obtain Ritz's approximate solution of the problem.

Let $b_1, b_2, \ldots, b_n$ be any suitable constants. We put

$$\zeta_n = b_1\psi_1 + b_2\psi_2 + \ldots + b_n\psi_n.$$

Multiplying equation (17) by b_i, and summing over all i we obtain, using formula (18),

$$\int_\Omega\int (\nabla^2 w_n \nabla^2 \zeta_n - f\zeta_n)\,dx\,dy = 0. \tag{19}$$

The solution of equations (17) when substituted in the integral J_n give it a minimal value which we shall denote by $J_n^{(0)}$. It is readily proved that

$$J_n^{(0)} = -\frac{1}{2}\int_\Omega\int (\nabla^2 w_n)^2\,dx\,dy. \tag{19a}$$

With increasing n the magnitude of $J_n^{(0)}$ goes down or, at least, does not increase (cf. § 14); at the same time this quantity is bounded below since it is not less than the exact lower bound of integral (11). According to a well-known theorem on a monotonic variable the quantity $J_n^{(0)}$ tends to some limit. On the basis of Cauchy's test for the existence of a limit (cf. 1, 42) for a given positive quantity η (no matter how small) it is possible to find a number N such that for $n > N$ and for arbitrary m the inequality

$$0 \leq J_n^{(0)} - J_{n+m}^{(0)} < \frac{1}{2}\eta \tag{20}$$

is satisfied. We shall use the notation $(w_{m+n} - w_n)/\sqrt{\eta} = \varphi(x, y)$. By using identity (19) in which n is replaced by $n+m$ and by performing various identity transformations we reduce inequality (20) to the form

$$\int_\Omega\int (\nabla^2\varphi)^2\,dx\,dy < 1. \tag{21}$$

To the functions $\varphi(x, y)$ we apply the well-known formula

$$\varphi(x,y)=\frac{1}{2\pi}\int_S\left(\varphi\frac{\partial\ln r}{\partial n}-\ln r\frac{\partial\varphi}{\partial n}\right)dS+\frac{1}{2\pi}\int\int_\Omega\nabla^2\varphi\cdot\ln r\,d\xi d\eta.$$

(See e.g., 2, **193**, formula (11).)

As is readily seen $\varphi(x, y)$ is a linear combination of the coordinate functions and thus satisfies conditions (10). Consequently the contour integral in the last formula vanishes and we obtain

$$\varphi(x,\ y)=\frac{1}{2\pi}\int\int_\Omega\nabla^2\varphi\ln r\ d\xi d\eta. \tag{22}$$

We now apply Cauchy's inequality (cf. § 3) to integral (22), whence it follows that:

$$|\varphi(x,\ y)|\leq\frac{1}{2\pi}\left\{\int\int_\Omega(\nabla^2\varphi)^2d\xi d\eta\right\}^{\frac{1}{2}}\left\{\int\int_\Omega\ln^2r\,d\xi d\eta.\right\}^{\frac{1}{2}}.$$

By inequality (21) the first integral is less than unity, the second, as is easily proved (although we shall not do so), does not exceed some constant number. But then $|(\varphi\ (x, y)|$ does not exceed some constant number; denoting this number by the letter K we have

$$|\varphi(x,\ y)|\leq K.$$

Now

$$|w_{m+n}(x,\ y)-w_n(x,\ y)|\leq K\gamma(\eta),\qquad n>N;$$

on the basis of Cauchy's test (cf. **1, 42**) the approximate solutions converge to some limiting function uniformly in the closed domain $\Omega+\mathrm{S}$

(An aggregate of points consisting of the points of a domain and the points on its boundary is called a closed domain.) The limiting function which we shall denote by $w(x,\ y)$ is continuous in $\Omega+\mathrm{S}$ as the limit of a uniformly convergent sequence of the continuous functions w_n (cf. **1, 145**) and equals zero on the boundary S.

Ritz notes that it is not possible to assert the uniform convergence of the derivatives of w_n; nevertheless he proves that the limiting function $w(x, y)$ possesses derivatives up to the fourth order inclusive within the domain Ω and satisfies the Sophie Germain equation. We have already seen that $w(x, y)$ satisfies the first of boundary conditions (10); in the paper mentioned Ritz tries to prove that the function $w(x, y)$ also satisfies the second of boundary conditions (10); however his reasoning is incorrect. The later reasoning of R. Courant, K. Friedrichs and S. L. Sobolev established that under general formulation of the equilibrium problem for a plate it is not possible to require that the condition

$$\left.\frac{\partial w}{\partial n}\right|_s = 0$$

be satisfied in the usual sense; however this condition is satisfied in a certain generalized sense (cf. R. Courant and D. Hilbert [2] and S. L. Sobolev [2]).

We now give brief mention to the solution of the Dirichlet problem given by Ritz. He considered the Poisson equation

$$\nabla^2 w = f(x, y), \tag{23}$$

which was to be integrated in the finite domain Ω of the (x, y) plane under the condition that on the boundary S of this domain the equation

$$w|_S = 0 \tag{24}$$

is satisfied. By again applying his method Ritz constructs a sequence of approximate solutions w_n, $n = 1, 2, \ldots$. But this time in the general case it is not possible to assert that this sequence converges uniformly. He merely succeeds in proving the existence of a function w such that the integrals

$$\int_\alpha^x w_m(\xi, y)\mathrm{d}\xi, \qquad \int_\beta^y w_m(x, \eta)\mathrm{d}\eta$$

converge uniformly to the integrals

$$\int_{a}^{x} w(\xi, y)\mathrm{d}\xi, \qquad \int_{\beta}^{y} w(x, \eta)\mathrm{d}\eta.$$

Moreover it may be proved that the function $w(x, y)$ satisfies equation (23). With respect to boundary condition (24) Ritz made a mistake analogous to that mentioned above.

In the works of Friedrichs, Rayleigh and Courant the energy method and the problem of its substantiation has received lengthy treatment. These authors studied variational problems corresponding to equations of more general form; they replaced the study of the convergence of the Ritz approximate solutions by the study of the more general problem of the convergence of the *minimizing sequence* (see below, § 13). In the works of the authors mentioned a strict formulation and a fairly complete solution of the problem of the eigenvalues of the Laplacian operator and of some others related to it are given. Many of the results due to the authors mentioned are given in a book by R. Courant and D. Hilbert [1, 2].

The well-known Parseval equations are of great importance in the substantiation of variational methods. These equations and their applications to mathematical physics are given by S. L. Sobolev in his monograph [2].

The author of the present book has given a representation of the solution of the variational problems, to which the usual problems of mathematical physics reduce, in the form of a series and has established the connexion between the Ritz method and this representation; he has also given a substantiation of the variational method for several new classes of equations, in particular, for equations whose ellipticity is violated on the domain boundary; he has also investigated the conditions for the convergence of the principal derivatives in the Ritz method. The author's results are presented in the book "Direct Methods in Mathematical Physics", in the monograph [11] and in a series of papers mentioned in the literature references and at the end of the book.

Together with the energy method there was evolved the closely related "Galerkin method" or more precisely the "Bubnov–Galerkin method".

In 1913 I. G. Bubnov published a report on the work of S. P. Timoshenko who applied the Ritz method to the study of the stability of plates and beams. Bubnov points out in this report that the equations to which the Ritz method leads can be obtained without resorting to the construction of the potential energy of the system and without solving the variational problem. Noting that Timoshenko seeks the displacement in the form

$$W = a_1\varphi_1 + a_2\varphi_2 + a_3\varphi_3 + \ldots, \tag{25}$$

where $\varphi_1, \varphi_2, \ldots$ are given functions; $\alpha_1, \alpha_2, \ldots$ are coefficients which are to be determined, Bubnov writes ". . . . extremely simple solutions can also be obtained in the usual way, i.e., without resorting to a consideration of the energy of the system, provided that convergence of the series for W is sufficiently great. In this case if we simply substitute the expansion for W in the general differential expression for equilibrium, multiply the expression obtained by $\varphi_k dx dy$ and integrate over the entire volume of the body, then we obtain an equation relating the coefficient a_k with all the others provided that the functions φ are chosen so that

$$\int\int \varphi_n \varphi_k dx dy = 0 \qquad \text{for } n \neq k.$$

This equation is valid for almost all the problems in the work considered since the author usually selects the expression for W in the form of trigonometric series, where this condition is satisfied; if for some reason an expansion in completely rational functions is required they can be replaced by spherical functions. By writing down from the relations obtained as many equations as there are terms in the series which we wish to retain, and by equating the determinant of the factors for the coefficients to zero, we reduce the problem of finding the critical load to the determination of the least root of an

integral rational function whose highest power equals the number of terms retained." Bubnov further writes: "We note that the results thus obtained do not require the solution of the differential equilibrium equation ..... and will be identical to those found by Prof. Timoshenko." If the quotations of Bubnov are put in symbolic form we obtain the following. The solution of the problem of the stability usually reduces to the integration of a differential equation of the form

$$Lw-\lambda Mw=0 \tag{26}$$

where w is the displacement, L and M are certain differential operators and λ is an unknown numerical parameter. The displacement w satisfies not only differential equation (26) but also certain homogeneous boundary conditions. Thus, for example, if we are concerned with the stability of a plate which is rigidly fixed at its edge, then the displacement w must also satisfy conditions (10). For an arbitrarily chosen λ the problem posed has only the trivial solution $w=0$; we thus become concerned with finding the values of λ for which equation (26) has a solution which is not identically zero; the lowest of these λ will define the critical load at which the plate loses stability. As we shall see below (§ 31) under well-known conditions this lowest λ equals the minimum of the ratio

$$\frac{\int\int_{\Omega} wLw\,dx\,dy}{\int\int_{\Omega} wMw\,dx\,dy}; \tag{27}$$

by using the boundary conditions it is usually possible to represent ratio (27) in a rather different form, this being the method used both by Ritz and by Timoshenko. The minimum of ratio (27) can be sought by the Ritz method which leads to a system of equations of the form

$$\sum_{k=1}^{n}(A_{ik}-\lambda B_{ik})a_k=0, \qquad i=1,\ 2,\ \ldots,\ n, \tag{28}$$

where a_k are the coefficients in expression (25) in which we take n terms, A_{ik} and B_{ik} are certain numbers which depend in a known way upon the coordinate functions φ_k. For the approximate solution (25) to be different from identically zero it is necessary and sufficient that the numbers a_k satisfying system (28) are not all equal to zero; for this in its turn it is necessary and sufficient that the determinant of system (28) vanish, i.e.

$$\begin{vmatrix} A_{11}-\lambda B_{11}, & A_{12}-\lambda B_{12}, \ldots, & A_{1n}-\lambda B_{1n} \\ A_{21}-\lambda B_{21}, & A_{22}-\lambda B_{22}, \ldots, & A_{2n}-\lambda B_{2n} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots \\ A_{n1}-\lambda B_{n1}, & A_{n2}-\lambda B_{n2}, \ldots, & A_{nn}-\lambda B_{nn} \end{vmatrix} = 0. \qquad (29)$$

Obviously equation (29) is an algebraic equation of the n th power in λ; the lowest root of this equation gives an approximate value for the minimum of ratio (27). The procedure described here is essentially that used by Timoshenko in the work which was criticized by Bubnov. However, the latter pointed out that *system* (28) *can be obtained in another, simpler manner*; it suffices to substitute in equation (26) the approximate value of w given by (25) in which n terms are retained, to multiply the expression obtained by φ_k, to integrate over the domain of the plate (or the length of the rod if we are solving the stability problem for a rod, etc.) and to equate the result obtained to zero. As the condition for the applicability of his procedure Bubnov gave the requirement of orthogonality of the coordinate functions, which in fact is not necessary.

This observation by Bubnov gave rise to a new direct method, the development and wide application of which is closely associated with the name of B. G. Galerkin. In his work [1] published in 1915 Galerkin solves a number of problems on the equilibrium and stability of rods and plates. The method which he applied is identical in form to the method obtained by Bubnov in his review of Timoshenko's equation mentioned above; however Galerkin's essentially new contribution was the fact that he did not connect his

method with any variational problem, so that it could be applied to any differential (and not only differential) equation and it did not require the orthogonality of the coordinate functions.

As we have already said Galerkin published his paper in 1915. Thereafter, a large amount of work has been published in which the Bubnov–Galerkin method has been extensively applied to the practical solution of extremely diverse applied problems. Papers also appeared in which this method was applied, not unsuccessfully, to the solution of non-linear problems. However, the substantiation of the method proved to be a matter of some difficulty. Only 25 years after the publication of Galerkin's paper did Repman [1] (1920) substantiate the Bubnov–Galerkin method in application to Fredholm-type integral equations; in the same year Petrov [1] obtained an analogous result for a certain special ordinary fourth-order differential equation. Shortly afterwards Keldysh [1] generalized Petrov's result for the case of an ordinary fourth-order differential equation and also proved the convergence of the Bubnov–Galerkin method in the case of the simplest boundary value problem for the general second-order elliptic equation. The author [4, 8] has obtained a fairly general sufficient criterion for the convergence of the Bubnov–Galerkin method and has demonstrated the application of this criterion to a number of problems; in particular, he has shown that all the results enumerated above are included in his criterion. We should also mention the work of Pol'sky [1, 2] in which the generalization given by Petrov for the Bubnov–Galerkin method is investigated. Finally Krasnosel'sky [1, 2] has substantiated the Bubnov–Galerkin method for a certain class of non-linear equations. Krasnosel'sky's method was used by Vorovich in his work [1] on the theory of envelopes. We have devoted Chapter IX to the Bubnov–Galerkin method.

Of the other direct methods the method of least squares and especially the method of networks deserve attention. The method of least squares has been applied to the solution

of Dirichlet's problem by Piccone [1]; some of its further applications are given in the works of Krylov, Kravchuk *et al.* (For greater detail see M.F. Kravchuk [1].)

The general conditions for the convergence of the method of least squares and a number of its applications are given in the author's works [3, 6]. A detailed account of the method of least squares will be given in Chapter X.

The method of networks is of great value. An extensive literature is devoted to this topic and we shall not dwell upon it here in any detail. An account of the method of networks can be found in the books by Kantorovich and Krylov [1], Panov [1] *et al.*

Of the greatest importance is the question of the estimation of the error of an approximate solution obtained by some direct method or other. In this connexion the so called *method of orthogonal projections* worked out in papers by Zaremba [1, 2] and Weyl [2] is especially important (see also M. I. Vishik [1–3]). This method is another variational method and allows us to construct approximate solutions of boundary value problems. An important characteristic emerges from this: knowing two approximate solutions of a given problem, one of which is constructed by the energy method and the other by the method of orthogonal projections, it is possible to estimate the error of both approximate solutions. Other methods also exist which play the same role in the estimation of error as the method of orthogonal projections just mentioned; the method due to Trefftz [1] which received further development in the works of Birman [3, 5, 6] is an example of such a method. We also recall the Friedrichs transformation which serves the same purpose. The reader can find a fairly detailed account of this transformation in the book by Courant and Hilbert [1].

Everywhere in Chapter I–V we shall consider only real functions of one or several real variables with the exception of some distinct places where the contrary is specially mentioned.

CHAPTER I

THE OPERATORS OF MATHEMATICAL PHYSICS

§ 1. The formulation of the basic problems

The partial differential equations of mathematical physics fall into two large classes. The equations of the first class describe processes in which the quantities sought, which are of interest to the investigator, vary appreciably with time, so that the quantities sought are functions of the coordinates and of time. An extremely simple and important representative of this class is the wave equation (2, **171**), a particular case of which is the equation governing the vibrations of a stretched string. Another important example of an equation of this class is the equation of heat flow (2, **203**). We shall not deal with these equations in this book. The equations of the second class describe stationary phenomena in which quantities of interest to the investigator do not vary with time or, at least, are of negligible variation. The most common of all such equations are of the so called *elliptical type.** We now consider one of the simplest problems of mathematical physics.

Let some body be inhomogeneously heated so that its temperature, which we denote by T, is a function of position within the body. If there are no heat sources within the body then this function satisfies the Laplace equation with three independent variables (2, **112**)

$$\nabla^2 T = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} = 0. \tag{1}$$

* (See, e.g. I. G. Petrovsky [2])

If we wish to find the temperature distribution within the body then it is necessary to integrate this equation, i.e., it is necessary to find a function satisfying this equation. It is easily seen, however, that there is an infinite class of functions which satisfy the Laplace equation, while it is clear physically that in a given body, under given conditions, the temperature distribution is unique.

This means that in solving our problem we cannot start only from the Laplace equation; it is necessary to use further conditions. We can arrive at these conditions, for example, thus: we are able to measure directly the temperature at any point on the surface of the given body. Suppose that we have measured the temperature at many points, fairly densely distributed over the surface of the body. It is then possible in practice to ensure that the temperature at any point on the surface of the body is known — this also serves as the supplementary condition for integration of the Laplace equation. We denote by Ω the spatial region filled by the heated body, and its bounding surface by S.

Further let P denote an arbitrary point on the surface S and $f(P)$ — the known temperature of the body at the point P on its surface. Our supplementary condition can be written thus:

$$T=f(P) \text{ on } S,$$

or, or rather more concisely

$$T|_S=f(P). \tag{2}$$

It is now possible to formulate mathematically the problem of the temperature distribution as follows: *find a function $T(x, y, z,)$ which satisfies the Laplace equation within the region and assumes the given values on its boundary S*; in mathematical physics this problem is known as the *Dirichlet problem*. The condition that the function sought assumes specified values at S is known as a *boundary* or *end condition*,

because it refers to the behaviour of the function sought at the boundary of the region.

Boundary conditions can be of a type other than that of the Dirichlet problem. Thus, suppose that we know the temperature T_0 of the surrounding medium and heat passes into this medium through the boundary S of the body. According to the law of heat loss formulated by Newton, the heat flow through the boundary is proportional to the temperature difference between the external medium and the surface of the body. This law leads us to a boundary condition of the form

$$\frac{\partial T}{\partial n} + h(T - T_0) = 0 \text{ on } S,$$

where n is the external normal to the surface S and h is a coefficient. Taking $hT_0 = \varphi(P)$, it is possible to rewrite the boundary in the form

$$\left[\frac{\partial T}{\partial n} + hT\right]_S = \varphi(P). \tag{3}$$

The problem of integrating the Laplace equation under boundary condition (3) is often known as the *mixed problem.*

If in the left-hand side of condition (3) we take $h = 0$, then we obtain a new boundary condition

$$\left.\frac{\partial T}{\partial n}\right|_S = \varphi(P). \tag{4}$$

The problem of integrating the Laplace equation under this condition is usually called the *Neumann problem.* We arrived at it here from entirely formal considerations and it is uncertain whether the Neumann problem thus established can have a large significance in the theory of heat transfer. However problems of this type are of importance in connexion with many other physical phenomena. Thus, for example, the problem of the bending of elastic rods leads to the seeking

of a function of two variables, $\varphi(x, y)$, which satisfies the "two-dimensional" Laplace equation

$$\frac{\partial^2\varphi}{\partial x^2}+\frac{\partial^2\varphi}{\partial y^2}=0$$

in the region of the base of the rod, and at the boundary of this region satisfies the boundary condition

$$\frac{\partial\varphi}{\partial n}=y\cos(n, x)-x\cos(n, y).$$

The problems of the flexure of an elastic rod by a transverse force, and the problem of the flow (of an ideal fluid) around a solid body together with many others, reduce to the Neumann problem.

Thus we have explained that when considering the Laplace equation it is expedient to pose and solve three problems which differ only in the boundary conditions associated with the equation; these are the Dirichlet, Neumann and mixed problems. Sometimes they are known simply as the first, second and third boundary conditions.

The Laplace equation is the simplest and most important representative of the class of elliptic equations. In mathematical physics other equations of this class are also often encountered. Thus, if within an unevenly heated body heat sources of intensity $f(x, y, z)$ are distributed, then the temperature $T(x, y, z)$ of this body no longer satisfies the Laplace equation, but the so called *Poisson equation*

$$\nabla^2T=-\frac{1}{k}f(x, y, z), \tag{5}$$

where k is the thermal conductivity. If the body be non-homogeneous, so that k is not constant but a function of the coordinates, then equation (5) is replaced by the following:

$$\nabla\cdot\{k\nabla T\}=-f(x, y, z). \tag{6}$$

Equations (5) or (6) can be associated with a boundary condition of one of the types (2), (3) or (4). For these equations it is also possible to pose other boundary conditions but their value is not great we shall not deal with them here; we merely note the problem, encountered in practice, when the boundary of the region is divided into parts on each of which one of the conditions of the form (2), (3) or (4) is specified. The problem of integrating a differential equation under some boundary condition or other is known as a *boundary-value problem.*

In practice it is often required to integrate a differential equation of elliptical type in two or three independent variables which are usually the cartesian coordinates of a variable point in space. As we shall see below, neither the formulation of the basic problems, nor the methods for their solution considered in this book, depend essentially in any way upon the number of dimensions of the space. Therefore in order to simplify the exposition, we shall often refer to a space of m dimensions, understanding by this that m can equal two or three; in many cases, however, the function of interest depends upon only one independent variable, and then $m=1$.

We observe here that, from the theoretical point of view, the number m of spatial dimensions can be anything we please.

Equations of the second order of elliptical type are insufficient however for a study of some important stationary phenomena. Thus the theory of the deformation of a thin plate leads to an equation of the *fourth* order

$$\frac{\partial^4 w}{\partial x^4}+2\frac{\partial^4 w}{\partial x^2 \partial y^2}+\frac{\partial^4 w}{\partial y^4}=f(x, y).$$

For its integration the function sought $\omega(x, y)$ obeys boundary conditions which depend upon the method of fixing at the rim of the plate. These conditions will be considered in more detail in § 27.

In a number of cases it is necessary to resort to *systems* of

equations in partial derivatives. Thus in elasticity theory we are concerned with a system of three second-order equations in three independent variables; three second-order equations with variable coefficients but only in two independent variables define the displacement of a thin elastic shell. The number of such examples could easily be increased.

We do not intend to carry out here a detailed analysis of the boundary conditions associated with partial differential equations of higher orders or with systems of such equations, such an analysis would be laborious and at the same time scarcely justified. In the following we shall formulate the simplest boundary conditions which emerge in a natural way from the requirements of each distinct problem.

§ 2. Some ancillary concepts and formulae

DOMAIN. As we have been able to see from the material of § 1, the object of mathematical physics is the finding of unknown functions which satisfy a given (usually partial) differential equation and some boundary condition(s) associated with it. This function must be defined in some domain; this domain will be a surface if the function sought depends upon two independent variables, or a volume if the number of independent variables equals three; it degenerates into a curve if we are dealing with a function of only one independent variable; in this case the ends of the segment are not added to the domain. If the function sought were to depend upon four or more independent variables, then the domain in which this function is defined would have to be in a *hyperspace*. Thus if we are solving the problem of the torsion of a prism, or in the more general case, of a cylindrical rod, then the torsion function must be defined in a plane domain which is a normal section of the rod; this domain can be a circle, a rectangle, an annulus and so on, depending on the shape of the rod. In the problem of the flexing of a disc the function sought is

the displacement of the disc; the domain in which it must be defined is the part of the plane of zero displacement intercepted by the disc itself. If we are studying the temperature distribution in a non-uniformly heated body then the domain in which the function sought, the temperature at each point in the body, must be defined as that part of space filled by the heated body.

According to the general definition, a domain is a set of points in space which is characterized by two properties: (1) if some point P belongs to the domain, then all points sufficiently close to P also belong to the domain; (2) any two points of the domain can be joined by a line lying entirely within the domain. The first property is equivalent to saying that the domain consists only of *internal points*; the second property is that the domain is *connected*.

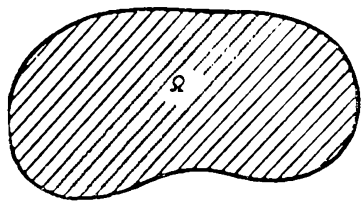

FIG. 1

The Boundary of a domain is defined as the set of points in any neighbourhood of each of which there are both points belonging to the domain and points not belonging to it. From property (1) of a domain it follows that pointson the boundary do not belong to the domain. The domain in which the function sought has to be defined we shall usually denote by the symbol Ω, the boundary of this domain by the symbol S. The boundary of a plane domain can be a closed curve (Fig. 1) or a set of several closed curves (Fig. 2.) similarly the boundary of a volume can consist of one or several closed surfaces. (We leave aside those cases in which the boundary of the region has a more complex structure.) For the purpose of the

present book, we are interested in those cases where the curve or surface forming the boundary of the region is either smooth, i.e. upon moving along such a curve (or surface) its tangent (or tangent plane) rotates continuously, or piecewise smooth, i.e. it consists of a finite number of smooth pieces. A circle or an ellipse serves as an example of a closed smooth curve and the boundary of a polygon or a semicircle as an example of a piecewise smooth curve. In the simplest case, where the function sought depends only upon one variable and the region degenerates into the segment of a line, the

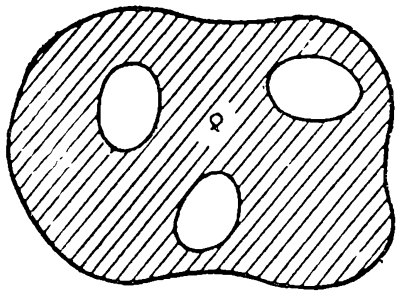

FIG. 2

boundary of the domain degenerates into a set of two points – the ends of this segment. We shall usually denote the points of the domain or of its boundary by the letters P or Q.

As has been stated, the points of the boundary are not part of the domain. The set of points which is obtained if we combine the domain and its boundary is known as a *closed domain*. It is usually denoted by the same letter as that used for the given domain, but with a bar above it, so that, for example $\overline{\Omega} = \Omega + S$. Sometimes a domain is known as an *open domain*, thereby distinguishing it from a closed domain.

In the following we shall consider only *finite domains*, i.e., domains which can be included within a sufficiently large sphere. The interior of a sphere, cube or torus can serve as an example of a finite domain and the exteriors of these bodies as examples of *infinite* regions.

The formula for integration by parts for multiple integrals is obtained as a direct consequence of Gauss's Theorem (2, **63**). For the case of a double integral it becomes Green's formula (2, **69**). Gauss's Theorem has the following form:

$$\int_{\Omega}\left(\frac{\partial\varphi}{\partial x}+\frac{\partial\psi}{\partial y}+\frac{\partial\omega}{\partial z}\right)\mathrm{d}\Omega$$

$$=\int_{S}(\varphi\cos(n,x)+\psi\cos(n,y)+\omega\cos(n,z))\mathrm{d}S;$$

if the region Ω is taken as three-dimensional. Here $\mathrm{d}S$ is an element of area of the surface S (the length of arc of the curve S, if the region Ω is a surface, n is the external normal to S. If ψ, φ, ω are considered as the components of some vector $\mathbf{w}$, then Gauss's Theorem can be written in the form

$$\int_{\Omega}\nabla\,.\,\mathbf{w}\mathrm{d}\Omega=\int_{S}\omega_n\mathrm{d}S.$$

This we shall call the vector form of Gauss's Theorem; by w_n is denoted the projection of the vector $\mathbf{w}$ upon the direction of the external normal to S. Let $\psi=\omega=0$, $\varphi=uv$. Differentiating in the volume integral and carrying one of the terms to the right-hand side, we obtain the formula for integration by parts:

$$\int_{\Omega}u\frac{\partial v}{\partial x}\mathrm{d}\Omega=-\int_{\Omega}v\frac{\partial u}{\partial x}\mathrm{d}\Omega+\int_{S}uv\cos(n,x)\mathrm{d}S.$$

If the domain Ω is not necessarily three-dimensional but more generally m-dimensional where m can be both larger and smaller than three, and if the cartesian coordinates are denoted by $x_1, x_2, \ldots, x_m$, then the formula for integration by parts can be written, in the general case, as

$$\int_{\Omega}u\frac{\partial v}{\partial x_i}\mathrm{d}\Omega=-\int_{\Omega}v\frac{\partial u}{\partial x_i}\mathrm{d}\Omega+\int_{S}uv\cos(n,x_i)\mathrm{d}S. \qquad (1)$$

We now apply the formula for integration by parts to the derivation of the so called *Green's formula* which plays a large role in mathematical physics. Let $u(P)$ and $v(P)$ be functions which are continuous, together with their first and second derivatives in the closed domain Ω, and arbitrary elsewhere. Further, let $A_{ik}(P)$, $i, k=1, 2, \ldots, m$ be given functions which together with their derivatives are continuous in that same closed domain; by the letter m we have, as usual, denoted the number of dimensions of the space. We further take it that

$$A_{ik}(P) = A_{ki}(P)$$

for any i and k. We consider the differential expression

$$Lu = -\sum_{i=1}^{m}\sum_{k=1}^{m}\frac{\partial}{\partial x_i}\left(A_{ik}(P)\frac{\partial u}{\partial x_k}\right)+C(P)u,$$

where $C(P)$ is a continuous function in the closed domain Ω. The reason for which we consider it convenient to take the minus sign in the expression for Lu will be explained in § 24. We shall denote this more concisely in the form

$$Lu = -\sum_{i,\,k=1}^{m}\frac{\partial}{\partial x_i}\left(A_{ik}\frac{\partial u}{\partial x_k}\right)+Cu. \tag{2}$$

We form the integral

$$\int_{\Omega} vLu\,\mathrm{d}\Omega = -\sum_{i,\,k=1}^{m}\int_{\Omega} v\frac{\partial}{\partial x_i}\left(A_{ik}\frac{\partial u}{\partial x_k}\right)\mathrm{d}\Omega+\int_{\Omega} Cuv\mathrm{d}\Omega. \tag{3}$$

In formula (1) we replace u by v and v by $A_{ik}\dfrac{\partial u}{\delta x_k}$. We then obtain a formula which we shall call *Green's first formula:*

$$\int_{\Omega} vLu\,\mathrm{d}\Omega = \int_{\Omega}\sum_{i,\,k=1}^{m}A_{ik}\frac{\partial u}{\partial x_k}\frac{\partial v}{\partial x_i}\,\mathrm{d}\Omega+\int_{\Omega} Cuv\mathrm{d}\Omega -$$

$$-\int_S v \sum_{i,k=1}^{m} A_{ik} \frac{\partial u}{\partial x_k} \cos(n, x_i) \mathrm{d}S. \tag{4}$$

Putting $v(P) \equiv u(P)$, we obtain *Green's second formula:*

$$\int_\Omega uLu \, \mathrm{d}\Omega = \int_\Omega \sum_{i,k=1}^{m} A_{ik} \frac{\partial u}{\partial x_i} \frac{\partial u}{\partial x_k} \mathrm{d}\Omega$$

$$-\int_S u \sum_{i,k=1}^{m} A_{ik} \frac{\partial u}{\partial x_k} \cos(n, x_i) \mathrm{d}S. \tag{5}$$

In Green's first formula we interchange u and v and subtract the result obtained from (4):

$$\int_\Omega (vLu - uLv) \mathrm{d}\Omega = \int_\Omega \left\{ \sum_{i,k=1}^{m} A_{ik} \left(\frac{\partial u}{\partial x_k} \frac{\partial v}{\partial x_i} - \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_k} \right) \right\} \mathrm{d}\Omega$$

$$-\int_S \left\{ v \sum_{i,k=1}^{m} A_{ik} \frac{\partial u}{\partial x_k} \cos(n, x_i) - u \sum_{i,k=1}^{m} A_{ik} \frac{\partial v}{\partial x_k} \cos(n, x_i) \right\} \mathrm{d}S.$$

We shall prove that the volume integral equals zero. In the second summation under the integral sign we shall write i in place of k and k in place of i, which is clearly permissible. Then the function to be integrated assumes the form

$$\sum_{i,k=1}^{m} (A_{ik} - A_{ki}) \frac{\partial u}{\partial x_k} \frac{\partial v}{\partial x_i} \equiv 0,$$

because $A_{ki} \equiv A_{ik}$. Now denoting for brevity

$$\sum_{i,k=1}^{m} A_{ik} \frac{\partial u}{\partial x_k} \cos(n, x_i) = Nu, \tag{6}$$

we obtain *Green's third formula*

$$\int_\Omega (vLu - uLv) \mathrm{d}\Omega = \int_S (uNv - vNu) \mathrm{d}S. \tag{7}$$

We consider a particular case. Let

$$A_{ii}=1, \quad A_{ik}=0, \quad i\neq k; \quad C=0.$$

Then $Lu=-\nabla^2 u$ ($\nabla^2$is the Laplacian operator) and $Nu=\dfrac{\partial u}{\partial n}$. We obtain in this fashion *Green's formulae for the Laplacian operator.*

$$-\int_{\Omega} v\nabla^2 u\,d\Omega=\int_{\Omega}\sum_{i=1}^{m}\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i}d\Omega-\int_{S} v\frac{\partial u}{\partial n}dS; \tag{8}$$

$$-\int_{\Omega} u\nabla^2 u\,d\Omega=\int_{\Omega}\sum_{i=1}^{m}\left(\frac{\partial u}{\partial x_i}\right)^2 d\Omega-\int_{S} u\frac{\partial u}{\partial n}dS; \tag{9}$$

$$\int_{\Omega}(v\nabla^2 u-u\nabla^2 v)d\Omega=\int_{S}\left(v\frac{\partial u}{\partial n}-u\frac{\partial v}{\partial n}\right)dS. \tag{10}$$

Similar formulae, which are also usually known as Green's formulae can also be constructed for differential operators of order higher than the second. We shall derive such formulae later as they are required.

§ 3. The scalar product of functions

In a given domain we need to consider not only the function sought, but also other functions – for example the term in elasticity theory problems determined by the nature of the externally applied stresses, the potentials from which approximate solutions of our problems are constructed, the approximate solutions themselves and many others. Upon these functions we must perform several simple operations from which we specially select the operation of *scalar multiplication* of two functions. We shall assume for the time being that the function considered is continuous in the closed domain $\bar{\Omega}=\Omega+S$.

Definition. The scalar product of two functions relative to a given domain is the integral of the product of those functions taken over the given domain. The operation of forming a scalar product is known as the scalar multiplication of the functions.

Thus, for example, the scalar product of the two functions $u(P)$ and $v(P)$ relative to the domain Ω is the integral

$$\int_{\Omega}\int u(P)v(P)\mathrm{d}x\mathrm{d}y,$$

if Ω is a surface, or

$$\int\int_{\Omega}\int u(P)v(P)\mathrm{d}x\mathrm{d}y\mathrm{d}z,$$

if the domain Ω is a volume. For brevity we shall, as in the previous paragraph, denote the integration over the domain Ω by one integral sign, and the element of area or volume (depending upon the number of dimensions of the field) by the symbol $\mathrm{d}\Omega$. The scalar product of the functions $u(P)$ and $v(P)$ is usually denoted by the symbol (u, v) so that

$$(u, v)=\int_{\Omega} u(P)v(P)\mathrm{d}\Omega. \tag{1}$$

Example 1. The domain Ω is the square $0<x<1$, $0<y<1$; $u(P)=x^2+y^2$, $v(P)=x+y$. In this case

$$(u,\ v)=(x^2+y^2,\ x+y)=\int_0^1\int_0^1 (x^2+y^2)(x+y)\mathrm{d}x\mathrm{d}y=\frac{5}{6}.$$

Example 2. The domain Ω is the sphere $x^2+y^2+z^2<1$; $u(P)=x^2$, $v(P)=1$.

$$(u, v)=(x^2,\ 1)=\int\int_{\Omega}\int x^2\mathrm{d}x\mathrm{d}y\mathrm{d}z.$$

To evaluate the integral we introduce spherical coordinates

$$x=r\sin\vartheta\cos\varphi,\quad y=r\sin\vartheta\sin\varphi,\quad z=r\cos\vartheta,$$

$$\mathrm{d}x\mathrm{d}y\mathrm{d}z=r^2\sin\vartheta\mathrm{d}r\mathrm{d}\vartheta\mathrm{d}\varphi;$$

$$(u, v)=\int_0^1 r^3\mathrm{d}r\int_0^{\pi}\sin^3\vartheta\mathrm{d}\vartheta\int_0^{2\pi}\cos^2\varphi\mathrm{d}\varphi=\frac{\pi}{3}.$$

Example 3. The domain Ω is the line segment $-1 < x < 1$; $u(P)=x$, $v(P)=3x^2-1$;

$$(u,v)=(x,3x^2-1)=\int_{-1}^{1} x(3x^2-1)\mathrm{d}x=0.$$

If in a given problem there is only one domain, so that there is no danger of ambiguity, then the words "relative to the given domain" in the definition of the scalar product are usually omitted and one speaks simply of the scalar product of two functions.

We note some simple, but extremely important properties of the scalar product, which we shall often use later.

A. *A scalar product is unchanged under interchange of the factors so that*

$$(u, v)=(v, u). \tag{2}$$

This property follows directly from formula (1).

B. *If each factor is the sum of several terms, then the scalar product can be expanded according to the usual rules of multiplication of polynomials and a constant factor can be placed outside the sign of the scalar product.*

Thus if a_1, a_2, b_1, b_2 are constants, then

$$\begin{aligned}(a_1u_1+a_2u_2,\ b_1v_1+b_2v_2)&=a_1b_1(u_1,\ v_1)\\&+a_1b_2(u_1,\ v_2)+a_2b_1(u_2,\ v_1)+a_2b_2(u_2,\ v_2).\end{aligned} \tag{3}$$

The proof is very simple. Indeed, by the definition of scalar product

$$\begin{aligned}&(a_1u_1+a_2u_2,\ b_1v_1+b_2v_2)\\&\quad=\int_{\Omega}[a_1u_1(P)+a_2u_2(P)]\ [b_1v_1(P)+b_2v_2(P)]\mathrm{d}\Omega.\end{aligned}$$

Removing the brackets under the integral sign and integrating term by term, we obtain

$$\begin{aligned}(a_1u_1+a_2u_2,\ b_1v_1+b_2v_2)&=a_1b_1\int_{\Omega}u_1(P)v_1(P)\mathrm{d}\Omega\\&+a_1b_2\int_{\Omega}u_1(P)v_2(P)\mathrm{d}\Omega+a_2b_1\int_{\Omega}u_2(P)v_1(P)\mathrm{d}\Omega\\&+a_2b_2\int_{\Omega}u_2(P)v_2(P)\mathrm{d}\Omega.\end{aligned}$$

By virtue of formula (1), the right-hand side of the last equation is identical with the right-hand side of formula (3) which is thus proved.

C. *The scalar product of any function with itself is non-negative.* In fact

$$(u, u) = \int_{\Omega} u^2(P) \mathrm{d}\Omega \geq 0. \tag{4}$$

D. *The scalar product of a function with itself reduces to zero if and only if this function is identically equal to zero.* This property of scalar multiplication immediately follows from relation (4).

The scalar product of functions is entirely similar to the well-known scalar product of vectors in vector analysis. This clearly possesses properties A–D. This analogy prompts us to introduce a further concept, analogous to the length of a vector. We recall that the length of a vector $\mathbf{A}$ can be expressed using the scalar product by the formula.

$$|\mathbf{A}| = \sqrt{(\mathbf{A} \cdot \mathbf{A})}.$$

We shall call the arithmetical square root of the scalar product of a function with itself, the *norm* of the function. The norm of the function $u(P)$ is usualiy denoted by the symbol $\|u\|$ so that

$$\|u\| = \sqrt{(u, u)} \tag{5}$$

or, by formula (4)

$$\|u\| = \sqrt{\left(\int_{\Omega} u^2(P) \, \mathrm{d}\Omega \right)}. \tag{6}$$

We now state the simplest properties of the norm; they follow readily from the definition of the norm and from properties A – D of the scalar product.

(a) *The norm of a function equals zero if and only if the function is identically equal to zero.* This property of the norm is a simple rephrasing of property D of the scalar product.

(b) $$\|au\| = |a| \|u\|; \qquad a = \text{const}. \tag{7}$$

In fact, by formula (3)

$$(au, au)=a^2(u, u);$$

and taking the root we arrive at formula (7).

Before formulating the next property of the norm, we recall that the scalar product of two vectors is defined by the formula $\mathbf{A}\cdot\mathbf{B}=|\mathbf{A}||\mathbf{B}|\cos\alpha$ where α is the angle between the vectors $\mathbf{A}$ and $\mathbf{B}$. Since $|\cos\alpha|\leq 1$ the scalar product of two vectors is related to their lengths by the inequality

$$|\mathbf{A}\cdot\mathbf{B}|\leq|\mathbf{A}||\mathbf{B}|.$$

A simple inequality occurs also for functions, i.e.:

(c) $$|(u, v)|\leq\|u\|\|v\|. \tag{8}$$

For the proof we take an arbitrary real number λ and construct the function $u(P)+\lambda v(P)$. By property D of the scalar product,

$$(u+\lambda v, u+\lambda v)\geq 0.$$

Removing the brackets according to formula (3), we reduce the latter inequality to the form

$$(u, u)+2\lambda(u, v)+\lambda^2(v, v)\geq 0; \tag{9}$$

where we have made use of property A, according to which $(u, v)=(v, u)$. The left-hand side of inequality (9) is quadratic in λ, and all values of it are non-negative. Hence it follows that its discriminant is less than or equal to zero:

$$(u, v)^2-(u, u)(v, v)\leq 0,$$

or, what is the same thing

$$|(u, v)|\leq\sqrt{(u, u)}\sqrt{(v, v)}.$$

The latter inequality, in virtue of the definition of the norm, is the same as inequality (8). If we recall the definitions of scalar product and norm, inequality (8) can be written in the more usual form

$$\left(\int_\Omega u(P)v(P)\,d\Omega\right)^2\leq\int_\Omega u^2(P)d\Omega\int_\Omega v^2(P)d\Omega. \tag{8_1}$$

Inequality (8_1) was first established by Bunyakovsky and bears his name. Even earlier the analogous inequality was obtained by Cauchy for finite and infinite sums. Cauchy's inequality has the form

$$(\sum a_k b_k)^2 \leq \sum a_k^2 \sum b_k^2. \tag{8_2}$$

We shall call inequality (8) the *Cauchy–Bunyakovsky inequality*. Example: Let Ω be the segment $0<x<1$, $u=x$, $v=x^2$. Then

$$(u, v)=\int_0^1 x^3 \mathrm{d}x=0{\cdot}25; \tag{9}$$

$$\|u\|=\sqrt{\left(\int_0^1 x^2 \mathrm{d}x\right)}=\frac{1}{\sqrt{3}};$$

$$\|v\|=\sqrt{\left(\int_0^1 x^4 \mathrm{d}x\right)}=\frac{1}{\sqrt{5}}$$

and clearly

$$0{\cdot}25<\frac{1}{\sqrt{3}}\frac{1}{\sqrt{5}}=0{\cdot}258\ldots.$$

Passing on to the fourth fundamental property of the norm, we note that it also has a simple analogue in vector analysis. Let two vectors $\mathbf{A}$ and $\mathbf{B}$ be given. Together with their sum $\mathbf{A}+\mathbf{B}$ they form a triangle (Fig. 3) and thus $|\mathbf{A}+\mathbf{B}| \leq |\mathbf{A}|+|\mathbf{B}|$. The norm also possesses an analogous property:

(d) *the norm of the sum of two functions does not exceed the sum of their norms*

$$\|u+v\| \leq \|u\|+\|v\|. \tag{10}$$

In order to prove this, consider the quantity

$$\|u+v\|^2=(u+v, u+v).$$

Expanding the brackets we obtain

$$\|u+v\|^2=(u, u)+2(u, v)+(v, v)=\|u\|^2+2(u, v)+\|v\|^2.$$

And substituting the Cauchy–Bunyakovsky inequality $(u,v)\leq\|u\|\|v\|$ in this, we obtain

$$\|u+v\|^2\leq\|u\|^2+2\|u\|\|v\|+\|v\|^2=(\|u\|+\|v\|)^2,$$

which is equivalent to relation (10).

The inequality (10) is known as the *triangle inequality.*

Example. Let Ω be the segment $0<x<1$, $u=x$, $v=x^2$, as before. As we have seen, $\|u\|=\frac{1}{\sqrt{3}}$, $\|v\|=\frac{1}{\sqrt{5}}$.

$$\|u+v\|=\sqrt{\left(\int_0^1(x+x^2)^2\mathrm{d}x\right)}=\frac{1}{30}\sqrt{930}=1{\cdot}0165\ldots$$

and, clearly

$$\|u+v\|<\|u\|+\|v\|=\frac{1}{\sqrt{3}}+\frac{1}{\sqrt{5}}=1{\cdot}0246\ldots.$$

In many cases the function considered is vectorial in nature. Thus, in problems of elasticity theory an important role is played by the displacement vector, and in hydrodynamic problems by the velocity of a moving particle.

In both cases we are dealing with vectors which depend upon the coordinates of a variable point or, more briefly, with a vector function of a point. Let $\mathbf{u}(P)$ and $\mathbf{v}(P)$ be two such vector functions, continuous in the closed domain $\bar{\Omega}$; we denote their components along the coordinate axes by u_x, u_y, u_z, and v_x, v_y, v_z respectively. The integral

$$(\mathbf{u},\mathbf{v})=\int_{\Omega}\mathbf{u}\cdot\mathbf{v}\mathrm{d}\Omega, \tag{11}$$

is known as the scalar product of the functions $\mathbf{u}(P)$ and $\mathbf{v}(P)$ where $\mathbf{u}\cdot\mathbf{v}$ is the "ordinary" scalar product of the vectors $\mathbf{u}$ and $\mathbf{v}$:

$$\mathbf{u}\cdot\mathbf{v}=u_xv_x+u_yv_y+u_zv_z,$$

so that, in more detail

$$(\mathbf{u},\mathbf{v})=\int_{\Omega}(u_xv_x+u_yv_y+u_zv_z)\mathrm{d}\Omega. \tag{11_1}$$

The norm of a vector function is defined, from formula (5) by the equality

$$\|\mathbf{u}\|^2=(\mathbf{u},\ \mathbf{u})=\int_{\Omega}(u_x^2+u_y^2+u_z^2)\mathrm{d}\Omega, \tag{12}$$

which can also be written in the form

$$\|\mathbf{u}\|^2=\int_{\Omega}|\mathbf{u}|^2\mathrm{d}\Omega. \tag{12_1}$$

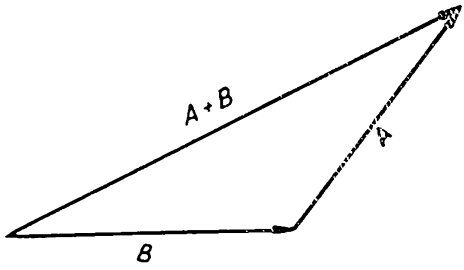

FIG. 3

The scalar product and norm in the case of vector functions possess properties A – D and (a) – (d), enumerated previously. We leave the proof of this to the reader.

In the foregoing we have assumed that the function considered is continuous in the closed domain. However, we often need to consider discontinuous functions; in particular we are interested in the class of functions continuous within the domain but discontinuous at its boundary, e.g. increasing infinitely upon approaching certain points on the boundary of the domain. Functions of interest are those which are continuous in the closed domain Ω except on certain points, lines or surfaces; in this case we shall consider discontinuities such that the integral [Understood in the Riemann sense as improper (**1, 116** and **2, 82**). A reader familiar with the fundamental theory of Lebesgue integrals can take it that we refer to functions which are, in the given region, quadratically integrable in the Lebesgue sense.] of the square of the function considered exists. If $u(P)$ and $v(P)$ are two such functions, so that the integrals

$$\int_{\Omega}u^2(P)\mathrm{d}\Omega, \qquad \int_{\Omega}v^2(P)\mathrm{d}\Omega$$

exist, then it may be proved that the integral

$$\int_{\Omega}u(P)v(P)\mathrm{d}\Omega$$

also exists.

For functions of this class we retain the definitions of scalar product and norm according to formulae (1) and (5). Properties A—D of scalar product and properties (a) — (d) of the norm in this case, obviously, are preserved.

§ 4. The concepts of operator and of functional

We wrote the equations of mathematical physics in § 1 so that there was a known function (in a particular case zero) on the right-hand side, and on the left-hand side an expression obtained from the function sought as the result of performing certain operations (addition, multiplication, differentiation etc.) the result of these operations upon the function sought again being some function.

For the solution of the problems of mathematical physics an investigation of the properties of the expression mentioned is particularly important; here it seems expedient not to restrict ourselves merely to the operations upon the function sought in a given problem, but to consider the results of the operations mentioned upon functions which are more or less arbitrary. This leads to the concept of an operator which is the basis of all that follows and which we shall define in the following way:

We shall say that *an operator is defined for some set of functions if a law is given according to which a new set of functions is defined such that to each function of the given set there corresponds one and only one function of the new set.*

As is easily seen the definition of an operator is virtually identical with the definition of a function (1, 5).

We shall sometimes say that an operator *acts* upon a function of the set for which it is defined. This set is known as the *field of definition* of the operator, and the new set of functions which the operator places in correspondence with the functions of its own field of definition is called its *field of values*. Functions belonging to the field of definition or to the field of values of some operator are often called *elements* of the corresponding field.

An operator is usually denoted by a symbol. Together with this is often written an indication of the function upon which the operator acts. Thus, we write, e.g. operator A or Au, where $u=u(P)$ is a function from the field of definition of the operator.

The field of definition of some operator A is customarily denoted by the symbol D_A, and its field of values by the symbol R_A.

An important example of an operator is the *Laplacian operator*

$$\nabla^2 u=\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}+\frac{\partial^2 u}{\partial z^2},$$

where x, y, z, are the cartesian coordinates of a point in three-dimensional space; in the general case of a space of m dimensions with cartesian coordinates $x_1, x_2, \ldots, x_m$ the Laplacian has the form

$$\nabla^2 u=\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}+\ldots+\frac{\partial^2 u}{\partial x_m^2}.$$

The field of definition of the Laplacian operator is the totality of doubly differentiable functions. In a number of cases, however, it is expedient to consider the Laplacian operator in a narrower field of definition. Thus, if for the Poisson equation

$$\nabla^2 u=f(x, y, z)$$

we seek a solution satisfying the boundary condition

$$u\big|_S=0,$$

then it is expedient from the very outset to consider the operator ∇^2 as acting not upon all possible functions but only upon those which are zero at the boundary of the domain.

Similarly, if the function sought must satisfy the boundary condition of the Neumann problem

$$\left.\frac{\partial u}{\partial n}\right|_S=0,$$

then we shall consider the field of definition of the Laplacian operator as the totality of those functions the normal derivative of which is zero on the boundary of the domain.

In the following we shall consider operators defined by different sets of functions as different operators even if the effects upon the functions are the same. Thus, the Laplacian operator defined, on the one hand for all functions which are only doubly differentiable, and on the other, for those functions which are doubly differentiable and, moreover, are zero on the boundary of the domain, will be considered as two distinct operators.

The Laplacian operator belongs to the class of *differential* operators, i. e. those containing the operation of differentiation. In mathematical physics we are also concerned with *integral* operators, containing the operation of integration. More rarely *integro-differential* operators are encountered. One of the simplest integral operators is the *Fredholm operator*

$$Ku = \int_{\Omega} K(P, Q)u(Q)\mathrm{d}\Omega_Q,$$

where $K(P, Q)$ is a given function of two points P and Q which satisfies certain requirements about regularity; e.g. it may be required that the function $K(P, Q)$, known as the kernel of the integral operator, be continuous when the points P and Q are within the closed domain $\bar{\Omega}$.

By enumerating these three classes, we have not, of course, exhausted the totality of all possible operators.

An operator which generates a function from itself is known as an *identity* operator. We shall denote it by the letter E, so that $Eu \equiv u(P)$ irrespective of the function $u(P)$. An operator which generates zero from any function is called a *null operator*. By the sum $A+B$, and the product AB of the operators A and B, are meant the operators defined by the formulae

$$(A+B)u = Au + Bu, \quad ABu = A(Bu).$$

The fields of definition of the operators $A+B$ and AB are the sets of functions for which the right-hand sides of the last

equations are meaningful. Obviously $A+B=B+A$, but, generally speaking $AB \neq BA$. If $AB=BA$ then the operators A and B are known as *permutable*. Let us write the statement "u is in the field of definition of A" $u \in D_A$. If the operators A and B are such that for any functions u and v such that $u \in D_A$ and $v \in D_B$ there occur the identities

$$BAu \equiv u, \quad ABv \equiv v,$$

then each of the operators A and B is called inverse with respect to the other. Usually in this case we write $B=A^{-1}$, $A=B^{-1}$. In the linear problems of mathematical physics, with which we shall be exclusively concerned here, an especially important role is played by the concept of the *linear* operator.

In order to define it we first of all introduce the concept of the *linear set* of functions.

A set of functions is called linear if it possesses the following property: if the functions u and v belong to the given set, then the functions $u+v$ and au, (where a is any constant), also belong to it.

The following are simple examples of linear sets of functions: (each set of functions is defined in one and the same domain) 1) The set of all functions, 2) The set of continuous functions, 3) The set of all polynomials, 4) The set of polynomials of degree not greater than a specified number, 5) The set of functions equal to zero at the boundary of the region. The reader can readily augment the number of such examples.

We now present some less simple examples: it happens that the set of functions, defined in a given domain Ω, the integrals of whose squares exist, is linear. Let $u(P)$ and $v(P)$ be such that the integrals

$$\int_\Omega u^2(P)\mathrm{d}\Omega, \quad \int_\Omega v^2(P)\mathrm{d}\Omega$$

exist. If a is a constant, then the same property is also possessed by the function $au(P)$, because

$$\int_\Omega (au)^2\mathrm{d}\Omega = a^2\int_\Omega u^2\mathrm{d}\Omega.$$

Further, from the inequality

$$(u+v)^2 \leq (u+v)^2 + (u-v)^2 = 2(u^2+v^2)$$

it follows that $\int_\Omega (u+v)^2 \, d\Omega$ exists, and that

$$\int_\Omega (u+v)^2 d\Omega \leq 2\int_\Omega u^2 d\Omega + 2\int_\Omega v^2 d\Omega.$$

We now cite some examples of non-linear sets. Of this type, for instance, is the set of all functions satisfying the inequality

$$|u(P)| < C, \tag{1}$$

where C is a positive constant. In fact, if the function $u(P)$ is different from zero at the point P_0, then for $a = \dfrac{2C}{u(P_0)}$ the function $au(P)$ no longer satisfies condition (1), since

$$|au(P_0)| = 2C > C.$$

A set of functions each of which is equal to unity at some point in the domain and zero elsewhere will serve as another example.

We note two obvious properties of a linear set: 1) any linear set of functions contains a function identically equal to zero; 2) if a linear set contains the functions $u_1(P)$, ... $u_n(P)$ and $a_1, a_2, \ldots a_n$ are arbitrary constants, then this set also contains the function

$$a_1u_1(P) + a_2u_2(P) + \ldots + a_nu_n(P).$$

We now give the definition of a linear operator. *The operator A is called linear if its field of definition is a linear set and if*

$$A(a_1u_1 + a_2u_2 + \ldots + a_nu_n) = a_1Au_1 + a_2Au_2 + \ldots + a_nAu_n, \tag{2}$$

irrespective of the integer n, $a_1, a_2, \ldots, a_n$ being constants and the functions $u_1, u_2, \ldots, u_n$ being from the field of definition of the operator A.

It is easily seen that any of the differential operators figuring in the left-hand sides of the equations of § 1 are linear when operating on a linear set; thus the Laplacian

operator is linear when specified by functions which have continuous first and second derivatives within a domain and are zero on the boundary of the domain. The identity operator and the null operator are also linear.

If we put $n=1$ and $a_1=0$ in equation (2) then we obtain $A0=0$. Thus any linear operator generates from a function identically zero the same function, identically equal to zero.

It may happen that from each function in its field of definition an operator generates a function identically equal to a constant. Operators of this kind are known as *functionals*. Thus, by definition, a functional generates from any function a constant. One of the simplest examples of a functional is the scalar product: if the function $v(P)$ is fixed, then the scalar product (u, v) is a functional on $u(P)$. Another simple example of a functional is the norm of a function.

In correspondence with the general definition of a linear operator the functional l is called *linear*, if its field of definition is linear and if

$$l(a_1u_1+a_2u_2+\ \ldots\ +a_nu_n)=a_1lu_1+a_2lu_2+\ \ldots\ +a_nlu_n,$$

where $a_1, a_2, \ldots, a_n$ are constants, and $u_1, u_2, \ldots, u_n$ are functions from the field of definition of the functional l. If the function v is fixed then the scalar product (u, v) is a linear functional with respect to u.

An important role will be played in the following by functionals of the form

$$F(u)=(Au, u)+lu+C, \tag{3}$$

where A is a linear operator, l is a linear functional and C is a constant. Such functionals are called *quadratic*.

§ 5. Symmetric operators

We now impose an essential restriction upon the operators that we shall study. When considering differential operators we shall require that the functions upon which they operate be

continuous within the closed domain and that the functions generated have a finite norm. For integral operators the first requirement can be weakened and we shall assume only that functions acted on by and generated by the integral operator have a finite norm.

By this requirement we restrict the class of functions upon which a given operator can act. Let, for example, Ω be the segment $0<x<1$ of the x-axis and A be the operator of double differentiation:

$$Au=\frac{\mathrm{d}^2u}{\mathrm{d}x^2}.$$

The function $u_1=\frac{1}{x}$ does not enter into the field of definition of our operator, since it is discontinuous at $x=0$. The function $u_2=x\ln x$ is continuous at $x=0$, however, it does not enter into the field of definition of our operator either, since the generated function $Au_2=\frac{\mathrm{d}^2u_2}{\mathrm{d}x^2}=\frac{1}{x}$ has an infinite norm. The function $u_3=x^2\ln x$ enters into the field of definition of the operator of double differentiation. In fact it is continuous, moreover the generated function $Au_3=\frac{\mathrm{d}^2u_3}{\mathrm{d}x^2}=2\ln x+3$, although discontinuous at $x=0$, has a finite norm:

$$\left\|\frac{\mathrm{d}^2u_3}{\mathrm{d}x^2}\right\|=\left\{\int_0^1(2\ln x+3)^2\mathrm{d}x\right\}^{\frac{1}{2}}=\sqrt{5}.$$

A linear operator A is called symmetric, if for any two functions u and v from its field of definition the following identity is valid.

$$(Au,\,v)=(u,\,Av). \tag{1}$$

The simplest example of a symmetric operator is the Friedholm operator

$$Ku=\int_{\Omega}K(P,Q)u(Q)\mathrm{d}\Omega_Q,$$

provided that its kernel satisfies the *symmetry condition*

$$K(P, Q)=K(Q, P).$$

In fact, writing $Ku=w$, we have

$$(Ku, v)=(w, v)=\int_{\Omega} v(P)w(P)\mathrm{d}\Omega_P$$

$$=\int_{\Omega} v(P)\,\mathrm{d}\Omega_P\int_{\Omega} K(P, Q)u(Q)\mathrm{d}\Omega_Q,$$

or, changing the order of integration,

$$(Ku, v)=\int_{\Omega} u(Q)\,\mathrm{d}\Omega_Q\int_{\Omega} K(P, Q)v(P)\mathrm{d}\Omega_P.$$

We replace the symbol P by Q and vice versa. This gives

$$(Ku, v)=\int_{\Omega} u(P)\mathrm{d}\Omega_P\int_{\Omega} K(Q, P)v(Q)\,\mathrm{d}\Omega_Q,$$

or, if we use the kernel symmetry condition,

$$(Ku, v)=\int_{\Omega} u(P)\mathrm{d}\Omega_P\int_{\Omega} K(P, Q)v(Q)\mathrm{d}\Omega_Q=(u, Kv).$$

We shall, however, generally be interested in symmetric *differential* operators; we now consider a few examples of such operators.

Let Ω be the segment $0<x<1$ and A be the operator of double differentiation, taken with a negative sign: $Au=-\dfrac{\mathrm{d}^2u}{\mathrm{d}x^2}$. Let the field of definition D_A of this operator be all the possible functions, continuous in the closed segment and having continuous first and second derivatives in this segment.

The operator A is asymmetric. In fact

$$(Au, v)-(u, Av)=-\int_0^1\left(v\frac{\mathrm{d}^2u}{\mathrm{d}x^2}-u\frac{\mathrm{d}^2v}{\mathrm{d}x^2}\right)\mathrm{d}x.$$

It is easily verified that

$$v\frac{d^2u}{dx^2}-u\frac{d^2v}{dx^2}=\frac{d}{dx}\left(v\frac{du}{dx}-u\frac{dv}{dx}\right);$$

whence

$$(Au, v)-(u, Av)=-[(vu'-uv')]_0^1.$$

In the general case the right-hand side of the last equality is different from zero and consequently $(Au, v)\neq(u, Av)$.

For example, if

$$u=x \text{ and } v=x^2, \text{ then } [(vu'-uv')]_0^1=-[x^2]_0^1=-1.$$

We now consider another example. Let B be the same operator of double differentiation with a negative sign, but with a more restricted field of definition: we assume that this field contains only functions which are such that they become zero at $x=0$ and at $x=1$. Elsewhere they are subject to the same conditions as in the previous example. If the functions u and v are in the field of definition of the operator B, then

$$u(0)=u(1)=v(0)=v(1)=0,$$

and

$$(Bu, v)-(u, Bv)=-\int_0^1\left(v\frac{d^2u}{dx^2}-u\frac{d^2v}{dx^2}\right)dx=-[(vu'-uv')]_0^1=0.$$

Hence it follows that the operator B is symmetric.

Let C mean the operator of double differentiation taken with a negative sign, the field of definition of which consists of functions subject to the boundary conditions

$$u'(0)+\alpha u(0)=0, \quad u'(1)+\beta u'(1)=0,$$

(where α and β are given numbers).

We now prove the symmetry of this operator. As before

$$(Cu, v)-(u, Cv)=-\int_0^1\left(v\frac{d^2u}{dx^2}-u\frac{d^2v}{dx^2}\right)dx=-[(vu'-uv')]_0^1,$$

or, at greater length

$$(Cu, v)-(u, Cv)=[u(1)v'(1)-v(1)u'(1)] \\ -[u(0)v'(0)-v(0)u'(0)]. \qquad (2)$$

We shall prove that the quantity in each square bracket on the right hand side equals zero. Both the functions $u(x)$ and $v(x)$ are in the field of definition of the operator C and thus satisfy the boundary conditions at $x=1$:

$$u'(1)+\beta u(1)=0, \quad v'(1)+\beta v(1)=0.$$

We multiply the first of these equations by $v(1)$ and the second by u (1). Subtracting, we obtain

$$u(1)v'(1)-v(1)u'(1)=0.$$

Similarly we may prove that the second bracket in (2) equals zero. Thus $(Cu, v)=(u, Cv)$ and the symmetry follows.

It is obvious that the identity operator and the null operator are symmetric.

It is easily seen that the sum of symmetric operators is also a symmetric operator; the product of symmetric operators, is symmetric if the operators permute. In fact, if A and B are symmetric operators, then applying formula (1) twice in succession, we obtain

$$(ABu, v)=(Bu, Av)=(u, BAv).$$

If A and B are permute then $AB=BA$. Substituting this in the last equation, we find

$$(ABu, v)=(u, ABv),$$

whence it follows that the product AB is symmetric.

We shall prove, finally that an operator inverse to a symmetric one is also symmetric. Let the symmetric operator A have the inverse operator A^{-1}. If $u \in D_A$ and $v \in D_A$ then, by formula (1)

$$(Au, v)=(u, Av).$$

Let f and g be arbitrary elements from the field of definition $D_{A^{-1}}$ of the inverse operator. If $u=A^{-1}f$ and $v=A^{-1}g$ then

$Au=f$, $Av=g$. Substituting this in the last equation we obtain

$$(f, A^{-1}g)=(A^{-1}f, g),$$

which proves the symmetry of the operator A^{-1}.

We now clarify this latter proposition by an example. We construct the operator reciprocal to the operator B of this section. In other words we shall find a function from the field of definition, D_B, of the operator B which satisfies the equation $Bu=f(x)$, where $f(x)$ is, say, an arbitrary, continuous function.

By the definition of the operator B the equation $Bu=f(x)$ can be written in the form

$$-\frac{\mathrm{d}^2u}{\mathrm{d}x^2}=f(x); \tag{3}$$

and the fact $u \in D_B$ means, moreover, that

$$u(0)=u(1)=0. \tag{4}$$

Thus, finding the quantity $B^{-1}f$ involves integrating equations (3) under boundary conditions (4). Integrating (3) twice, we obtain

$$u(x)=-\int_0^x \mathrm{d}s \int_0^s f(t)\mathrm{d}t+C_1x+C_2.$$

The condition $u(0)=0$ immediately gives $C_2=0$, and consequently

$$u(x)=-\int_0^x \mathrm{d}s \int_0^s f(t)\mathrm{d}t+C_1x.$$

We now change the order of integration, using Dirichlet's formula, (2, **72**)

$$u(x)=-\int_0^x f(t)\mathrm{d}t \int_t^x \mathrm{d}s + C_1x=-\int_0^x (x-t)f(t)\mathrm{d}t+C_1x.$$

From the second boundary condition $u(1)=0$ we find

$$C_1=\int_0^1 (1-t)f(t)\,\mathrm{d}t.$$

Writing this in the form

$$C_1=\int_0^x (1-t)f(t)\mathrm{d}t+\int_x^1 (1-t)f(t)\,\mathrm{d}t,$$

substituting this in the expression for $u(x)$ and combining the integrals which have the same limits, we find

$$u(x)=B^{-1}f=\int_0^x t(1-x)f(t)\mathrm{d}t+\int_x^1 x(1-t)f(t)\mathrm{d}t.$$

We now introduce the kernel $K(x,\ t)$ defined by the formula

$$K(x,\ t)=\begin{cases} x(1-t),\ x\leq t,\\ t(1-x),\ x\geq t.\end{cases}$$

This kernel is clearly symmetric. With its aid the operator B^{-1} can be represented in the form

$$B^{-1}f=\int_0^1 K(x,\ t)f(t)\mathrm{d}t.$$

Thus, B^{-1} is an integral operator with a symmetric kernel and is thus symmetric itself as it must be by virtue of the general proposition proved above.

The example just worked out has a further interest in that it shows the value of integral operators in mathematical physics: they arise as operators which are reciprocal to differential operators, in other words, they arise as a result of the solution of the differential equations of mathematical physics.

§ 6. Positive definite and positive-bounded-below operators

A symmetric operator A is said to be positive definite if for any function from its field of definition, $u(P)$ not identically equal to zero, the inequality

$$(Au,\ u)>0$$

is valid.

If $u(P)\equiv 0$, then clearly $(Au, u)=0$. Thus the definition given can be restated as follows. A symmetric operator A is positive definite if for any function $u(P)$ from its field of definition there occurs the inequality

$$(Au, u)\geq 0,$$

the equality $(Au, u)=0$ being possible only if $u(P)\equiv 0$.

As an example we consider the symmetric operator B of the previous section.

$$Bu=-\frac{\mathrm{d}^2u}{\mathrm{d}x^2},$$

the function $u(x)$ being subject to the boundary conditions

$$u(0)=u(1)=0; \tag{1}$$

the field Ω being the segment $0<x<1$. We now form the scalar product

$$(Bu, u)=-\int_0^1 u\,\frac{\mathrm{d}^2u}{\mathrm{d}x^2}\,\mathrm{d}x.$$

Integrating by parts, we have

$$(Bu, u)=+\int_0^1\left(\frac{\mathrm{d}u}{\mathrm{d}x}\right)^2\mathrm{d}x-[uu']_0^1\ .$$

In virtue of the boundary conditions (1) the last term vanishes, so that

$$(Bu, u)=\int_0^1\left(\frac{\mathrm{d}u}{\mathrm{d}x}\right)^2\mathrm{d}x\geq 0.$$

We now suppose that $(Bu, u)=0$. This means that

$$\int_0^1\left(\frac{\mathrm{d}u}{\mathrm{d}x}\right)^2\mathrm{d}x=0.$$

The integrand is non-negative and from the vanishing of the

integral it follows that $\frac{du}{dx}=0$ and, consequently, $u(x)=\text{const}$. Boundary condition (1) then shows that $u\equiv 0$. From what has been said it follows that the operator B is positive definite.

We consider another operator Cu, also considered in the previous section. We recall that functions from the field of definition of the operator C satisfy the boundary conditions:

$$u'(0)+\alpha u(0)=0, \qquad u'(1)+\beta u'(1)=0. \tag{2}$$

We have

$$(Cu,\ u)=-\int_0^1 u\,\frac{d^2u}{dx^2}\,dx=\int_0^1\left(\frac{du}{dx}\right)^2dx-[uu']_0^1$$
$$=\int_0^1\left(\frac{du}{dx}\right)^2dx-u(1)u'(1)+u(0)u'(0).$$

Eliminating $u'(0)$ and $u'(1)$ with the aid of the boundary condition (2), we obtain

$$(Cu,\ u)=\int_0^1\left(\frac{du}{dx}\right)^2dx+\beta u^2(1)-\alpha u^2(0).$$

It is difficult to judge whether the operator C is positive definite for arbitrary values of α and β.

It is easy however, to state simple sufficient conditions: the operator C is positive definite if $\alpha\leq 0$, $\beta\geq 0$ at least one of the constants α or β being different from zero. Consider, for example, $\beta\geq 0$ and $\alpha=-\gamma$ where $\gamma>0$. Then

$$(Cu,\ u)=\int_0^1\left(\frac{du}{dx}\right)^2dx+\beta u^2(1)+\gamma u^2(0)\geq 0.$$

If $(Cu,\ u)=0$ then it easily follows that, $u(0)=0$ and $\frac{du}{dx}\equiv 0$. From the last equality it follows that $u(x)=\text{const}$, and since $u(0)=0$ clearly $u(x)\equiv 0$.

If both constants α and β equal zero then the operator C is positive indefinite. In fact, in this case condition (2) assumes the form

$$u'(0)=u'(1)=0; \tag{3}$$

and

$$(Cu, u)=\int_0^1\left(\frac{\mathrm{d}u}{\mathrm{d}x}\right)^2\mathrm{d}x.$$

The function $u\equiv 1$ satisfies boundary conditions (3) and thus belongs to the field of definition of operator C. For this, $(Cu, u)=0$.

We now consider a more complex example. Let ∇^2 mean the Laplacian operator; the functions forming its field of definition are subject the boundary conditions

$$u\,|_S=0, \tag{4}$$

where S is the boundary of the domain Ω in which the function is defined.

We shall prove that in such a field of definition the operator $-\nabla^2$ is positive definite. We have

$$(-\nabla^2u, u)=-\int_\Omega u\nabla^2u\mathrm{d}\Omega=\int_\Omega\sum_{i=1}^m\left(\frac{\mathrm{d}u}{\mathrm{d}x_i}\right)^2\mathrm{d}\Omega-\int_S u\frac{\partial u}{\partial n}\mathrm{d}S;$$

where we have used formula (9) of § 2. By condition (4) the surface integral vanishes, and

$$(-\nabla^2u, u)=\int_\Omega\sum_{i=1}^m\left(\frac{\partial u}{\partial x_i}\right)^2\mathrm{d}\Omega\geq 0. \tag{5}$$

In this case, if $(-\nabla^2u, u)=0$ then of necessity $\dfrac{\partial u}{\partial x_i}\equiv 0$ since the integrand in the last integral is non-negative. But then $u=\mathrm{const}$ and in virtue of boundary condition (4), $u\equiv 0$.

We shall try to elucidate the physical meaning of the concept of a positive definite operator. To this end we con-

sider a membrane fixed at its edges, and bent under the action of some transverse load. In the unloaded state, let the membrane occupy some domain Ω on the plane (x, y) and let $u(x, y)$ be the transverse deflexion of the membrane. We denote by T the tension in the membrane, and by q the distorting load per unit area. As is known, the static deflexion of the membrane satisfies the differential equation

$$-\nabla^2 u = \frac{q}{T}; \tag{6}$$

the fact that the rim of the membrane is fixed is expressed by the boundary condition (4) in which S now denotes the edge of the membrane. (Equation (6) can easily be obtained from the equation of oscillation of the membrane (2, **176**) the displacement being independent of time. For the derivation of the equation of oscillation of the membrane see V. I. Levin and Yu. I. Grosberg [1] or A. N. Tikhonov and A. A. Samarsky [1]. The equation of the static displacement of a membrane is also derived in S. P. Timoshenko [2].)

On the other hand, in the given case the expression

$$(-\nabla^2 u, u) = \int_\Omega \int \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right] dx dy$$

only differs by a constant factor from the potential energy of deformation of the bent membrane. (See e. g. R. Courant and D. Hilbert [1] Ch. IV, § 10, p. 238.) Now it is clear that in the problem of the static deflection of a membrane the positive definiteness of the operator expresses the physically obvious fact that the potential energy of deformation of the membrane, however it is bent, is positive; in other words that it is impossible to distort the membrane without expending energy.

Many phenomena (some of which will be considered in the course of this book) possess the following property. Some physical system under the action of an external influence,

characterized by some function $f(P)$, acquires a displacement $u(P)$ these two functions being connected by the equation

$$Au=f,$$

where A is a positive definite operator. Then, it often happens, *that the quantity (Au, u) is proportional to the amount of energy which must be expended in giving a displacement $u(P)$ to the system.*

If this be so, then the positive definiteness of the operator expresses the fact that it is impossible to give any displacement to the system without expending energy.

If the operator A is positive definite, then bearing in mind the physical meaning of the quantity (Au, u) just stated, we shall further call this quantity the *energy of the function* $u \in D_A$.

A symmetric operator A is said to be positive-bounded-below if for any function $u \in D_A$ the inequality

$$(Au, u) \geq \gamma^2 \|u\|^2, \tag{7}$$

is valid, where γ is a positive constant.

Obviously any positive-bounded-below operator is, at the same time, positive definite. The converse conclusion, as we shall see by an example, is incorrect.

Example. Consider the operator

$$Bu=-\frac{d^2u}{dx^2},$$

where the functions $u(x)$ are defined in $0 \leq x \leq 1$ and satisfy the boundary conditions $u(0)=u(1)=0$. We saw previously that the operator B is positive definite. We shall prove that it is also positive-bounded-below.

Since $u(0)=0$,

$$u(x)=\int_0^x u'(t)dt.$$

We now apply Cauchy's inequality [formula (8_1) § 3] to the integral. This gives

$$u^2(x) \leq \int_0^x 1^2 dt \int_0^x u'^2(t)dt = x\int_0^x u'^2(t)dt.$$

Since $0<x<1$, we strengthen the latter inequality by changing the upper limit of the integral to unity. Thus

$$u^2(x)\leq x\int_0^1 u'^2(t)\mathrm{d}t.$$

Integrating this inequality over the range $0<x<1$, we find

$$\int_0^1 u^2(x)\mathrm{d}x\leq \frac{1}{2}\int_0^1 u'^2(t)\mathrm{d}t. \tag{8}$$

We now recall that by the definition of the norm

$$\int_0^1 u^2(x)\mathrm{d}x=\|u\|^2,$$

and for the operator B considered in our example

$$\int_0^1 u'^2(t)\mathrm{d}t=(Bu,\ u).$$

Inequality (8) now gives

$$(Bu,\ u)\geq 2\|u\|^2.$$

Thus, operator B is positive-bounded-below, and for this operator it is possible to take $\gamma=\gamma(2)$. Further on it will be shown that in many important cases the operators of mathematical physics are positive-bounded-below.

We now consider a further example. Let the operator L be specified by the formula

$$Lu=-\frac{\mathrm{d}}{\mathrm{d}x}\left(x^3\frac{\mathrm{d}u}{\mathrm{d}x}\right),\quad 0<x<1,$$

where we subject the function $u(x)\in D_L$ to the one boundary condition $u(1)=0$. It is easily seen that the operator L is positive definite. In fact

$$(Lu,\ v)-(u,\ Lv)=\int_0^1\left[u\frac{\mathrm{d}}{\mathrm{d}x}\left(x^3\frac{\mathrm{d}v}{\mathrm{d}x}\right)-v\frac{\mathrm{d}}{\mathrm{d}x}\left(x^3\frac{\mathrm{d}u}{\mathrm{d}x}\right)\right]\mathrm{d}x$$

$$=\int_0^1\frac{\mathrm{d}}{\mathrm{d}x}\left[x^3\left(u\frac{\mathrm{d}v}{\mathrm{d}x}-v\frac{\mathrm{d}u}{\mathrm{d}x}\right)\right]\mathrm{d}x=\left[x^3\left(u\frac{\mathrm{d}v}{\mathrm{d}x}-v\frac{\mathrm{d}u}{\mathrm{d}x}\right)\right]_0^1=0,$$

since the factor x^3 vanishes at $x=0$ and the functions $u(x)$ and $v(x)$ vanish at $x=1$. Now $(Lu,\ u)=(u,\ Lu)$ and the operator is symmetric.

Further

$$(Lu,u)=-\int_0^1 u\frac{\mathrm{d}}{\mathrm{d}x}\left(x^3\frac{\mathrm{d}u}{\mathrm{d}x}\right)\mathrm{d}x.$$

Integrating by parts and using the condition $u(1)=0$, we find that

$$(Lu, u)=\int_0^1 x^3\left(\frac{du}{dx}\right)^2 dx\geq 0.$$

In this case, if $(Lu, u)=0$ then $\frac{du}{dx}\equiv 0$, $u=\text{const}$ and since $u(1)=0$, $u\equiv 0$. Thus the operator L is positive definite.

We shall now prove that this operator is not positive-bounded-below. We write inequality (7) in the form

$$\frac{(Au, u)}{\|u\|^2}\geq\gamma^2.$$

Hence it is clear that for a positive-bounded-below operator the latter ratio does not become less than the fixed positive constant γ^2. If we establish that for the operator L a similar ratio can be made as small as we please, then it will follow that the operator L is not positive-bounded-below.

Let δ be a number included between zero and unity. We assume

$$u_\delta(x)=\begin{cases}(\delta-x)^3, & 0\leq x\leq\delta,\\ 0, & \delta<x\leq 1,\end{cases}$$

the function $u_\delta(x)$ equals zero at $x=1$ and is continuous, as are its first and second derivatives, for $0\leq x\leq 1$. Thus u_δ enters into the field of definition D_L of the operator L. We now form the ratio

$$\frac{(Lu_\delta, u_\delta)}{\|u_\delta\|^2}=\frac{\int_0^1 x^3\left(\frac{du_\delta}{dx}\right)^2 dx}{\int_0^1 u_\delta^2 dx}=\frac{9\int_0^\delta x^3(\delta-x)^4 dx}{\int_0^\delta(\delta-x)^6 dx},$$

since $u_\delta\equiv 0$ for $x>\delta$ and the corresponding integral vanishes. Performing a simple calculation we find

$$\frac{(Lu_\delta, u_\delta)}{\|u_\delta\|^2}=\frac{9\delta}{40},$$

which can be made as small as we please for sufficiently small δ.

We now give a physical explanation of an operator being positive-bounded-below. We shall compare the displacement at a point which the physical system experiences with the norm of the displacement. We shall term the displacement "large" if it is greater than the norm, and "small" if it is less than the norm. With this criterion relatively large local

displacements which do not seriously effect the norm are not of interest. Inequality (7) shows that for a system involving a positive-bounded-below operator, it is possible to obtain a large displacement only by a sufficiently great expenditure of energy. If the operator is positive definite, but not positive-bounded-below, it is possible to give the system an arbitrarily large displacement with the expenditure of an arbitrarily small amount of energy.

CHAPTER II

ENERGY CONVERGENCE

§ 7. The estimation of the error of an approximation and the nature of its convergence. Convergence in the mean

Direct methods furnish, as a rule, *approximate* solutions only of problems in which they are used. This occasions the necessity of estimating the error introduced by using, instead of the exact solution of the problem, the approximate solution that we have constructed. For example, let the function $u(P)$ – the exact solution – be replaced by another function $u_A(P)$ which is an approximate solution of the same problem. The error is $|u(P)-u_A(P)|$. The question arises of the estimation of this error. If we were concerned not with functions but with numbers the problem would readily be solved. In fact, if the problem consists of finding some number a and we find its approximate value b, then for an estimation of the error it suffices to estimate the magnitude of $|b-a|$ and we would consider as the best approximate method that which, with the same amount of effort, gives the least magnitude for $|b-a|$. The difference $|u(P)-u_A(P)|$ has different values at different points. Thus it is not obvious how to estimate its degree of approximation. In such cases, as a measure of the error we usually select some number related to the function $|u(P)-u_A(P)|$ in a defined fashion. It is simplest of all to take as a measure of the error the number $\varrho=\max |u(P)-u_A(P)|$. If the magnitude of ϱ is small then the difference in the values of the functions $u(P)$ and $u_A(P)$ will be small (it cannot exceed the small magnitude

ϱ) at any point in the region considered. In this case it is said that the function $u_A(P)$ is *uniformly close* to $u(P)$; the number ϱ can be called the measure of the closeness of both functions. Thus for small values of x the functions x and $\sin x$ are close to each other and $|\sin x - x| < \frac{|x|^3}{6}$. If, for example we restrict ourselves to values of $|x| \leq \frac{\pi}{6} = 0{\cdot}5236$ then, $|\sin x - x| < 0{\cdot}024$. Thus for any value of x between $-\frac{\pi}{6}$ and $+\frac{\pi}{6}$ the error in the approximate equality $\sin x \simeq x$ remains less than $0{\cdot}024$.

It is often important, not only for the functions themselves to be close to one another, but also their respective derivatives up to some order k. In such a case, as a measure of the closeness of the functions $u(P)$ and $u_A(P)$ (or what comes to the same thing, as a measure of the error in the approximate equality $u \simeq u_A$) it is appropriate to define,

$$\varrho^{(k)} = \max \left\{ |u(P) - u_A(P)| + \sum_{i \leq k} |D^{(i)}u - D^{(i)}u_A| \right\},$$

where $D^{(i)}$ means any derivative of order i. Under the sign of summation there enter all possible derivatives of order $\leq k$. If the number $\varrho^{(k)}$ is small then both the functions themselves and their derivatives up to and including order k differ one from one another by a uniformly small amount.

Other cases are encountered where the number ϱ is not the most convenient measure of closeness. It frequently happens that the difference in the functions $u_A(P)$ and $u(P)$, while small in general, is significant over certain regions. In Fig. 4 are drawn the graphs of the functions $y = x$ and $y = x + 20xe^{-400x^2}$ over the range of the variable x from zero to unity. It is clear from the figure that these graphs diverge significantly only over the small section from $x = 0$ to $x = 0{\cdot}13$; for $x > 0{\cdot}13$ the graphs coincide. At the same time the

maximum separation of the graph is considerable; it equals 0·43 and occurs at $x=\frac{\gamma(2)}{40}\simeq 0\cdot 035$. From the point of view of uniform closeness the functions x and $x+20xe^{-400x^2}$ are essentially different, while intuition would suggest that this

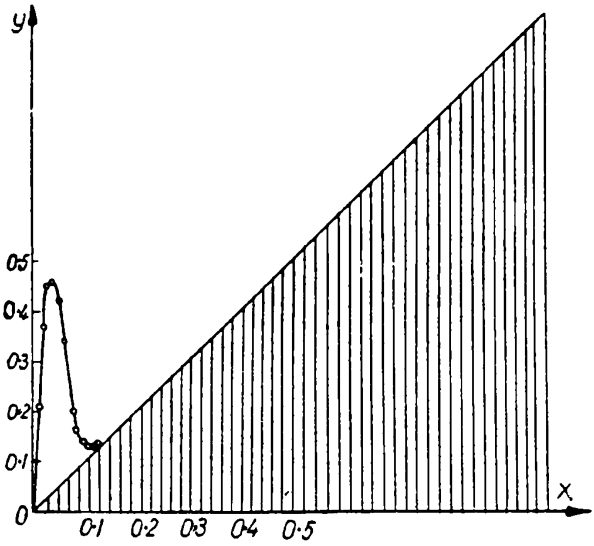

FIG. 4

is not entirely so and that if we were interested, for example, in the area between the two curves, then we could easily satisfy ourselves that, from this point of view, the graphs differ very little. In fact, the difference in the areas below the two curves equals

$$\int_0^1 20xe^{-400x^2}dx.$$

Putting $400x^2=t$, we find that the area mentioned equals

$$\frac{1}{40}\int_0^{400} e^{-t}dt<\frac{1}{40}\int_0^{\infty} e^{-t}dt=0\cdot 025,$$

which constitutes, in all, 5% of the hatched triangle in Fig. 4. If we were to compare the graphs of the functions x and $100xe^{-10000x^2}+x$ then it would happen that the maximum difference in the ordinates would be as before, 0·43, but a significant separation of the graphs would occur only over the range $0<x<0{\cdot}03$ and the difference in area would be less than 1% of the area of the hatched triangle. It is said of such functions that they are *close in the mean*; as a measure of closeness in the mean of the two functions $u(P)$ and $u_A(P)$ we often adopt the quantity

$$\delta=\left\{\int_\Omega [u(P)-u_A(P)]^2 \mathrm{d}\Omega\right\}^{\frac{1}{2}}=\|u-u_A\|.$$

It is sometimes necessary not only that the functions themselves be close in the mean but also their derivatives up to a certain given order k. Then it is expedient to take as a measure of the closeness the quantity

$$\delta^{(k)}=\left\{\int_\Omega [u(P)-u_A(P)]^2 \mathrm{d}\Omega+\sum_{i\leq k}\int_\Omega [D^{(i)}u-D^{(i)}u_A]^2 \mathrm{d}\Omega\right\}^{\frac{1}{2}}.$$

We shall not dwell here upon other possible methods of estimating the closeness of two functions; one such method will however be investigated in detail in the next paragraph.

The measure of the closeness of two functions is often called the *distance* between them, so that it is possible, for instance, to speak of the distance between the given functions $u(P)$ and $u_A(P)$ in the sense of uniform closeness; this is $\max |u(P)-u_A(P)|$. It is also possible to speak of the distance between these functions in the sense of closeness in the mean; this distance is $\|u-u_A\|$.

Closely connected with the concept of estimating the closeness of functions is the concept of the *convergence* (or of *tending to a limit*) of a sequence of functions. Let the distance between functions be defined in some way. Let some sequence of functions $u_1(P)$, $u_2(P)$, ..., $u_n(P)$, ... be given and let there exist a function $u(P)$ such that the distance between

the functions $u_n(P)$ and $u(P)$ tends to zero as $n \to \infty$. Then it may be said that $u_n(P)$ *converges* (or *tends*) *to* $u(P)$ *in a sense corresponding to the sense of the closeness.* Thus let the functions $u_n(P)$ and $u(P)$ be defined in the closed domain $\overline{\Omega}$. If $\max_{P \in \overline{\Omega}} |u_n(P) - u(P)| \to 0$ then $u_n(P) \to u(P)$ *uniformly* in $\overline{\Omega}$.

If $u_n(P)$ and $u(P)$ are defined in the domain Ω and $\|u_n - u\| \to 0$, then we come to a new type of convergence; in this case it is said that $u_n(P) \to u(P)$ in the *mean* in the domain Ω.

It is also possible to introduce, together with the concept of the convergence of a sequence, the concept of *the convergence of a series in a sense corresponding to the sense of the closeness*; the series

$$\sum_{k=1}^{\infty} u_k(P)$$

converges in some sense or other and has as the sum the function $u(P)$ if the sequence of partial sums of this series converges in the same sense to the function $u(P)$. In particular it is possible to speak of uniform convergence of a series and of it being convergent in the mean. We shall consider the concept of uniform convergence to be sufficiently well known to the reader (cf. 1 ch IV, § 3). In order to clarify the important concept of convergence in the mean, we now present an example of it. Let Ω be the interval $0 < x < 2\pi$ and let $f(x)$ be some absolutely integrable function. We construct its Fourier series

$$f(x) \sim \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx); \qquad (1)$$

and we denote the partial sums of the Fourier series by $S_n(x)$. As is well known, series (1) can, for some values of x, diverge. Nevertheless, we shall now satisfy ourselves that

this series always converges in the mean to $f(x)$, provided that this function has a finite norm. In fact

$$\|f(x)-S_n(x)\|^2=\int_0^{2\pi}\left\{f(x)-\frac{a_0}{2}-\sum_{k=1}^{n}(a_k\cos kx+b_k\sin kx)\right\}^2 dx$$

$$=\int_0^{2\pi}f^2(x)dx-\frac{\pi a_0^2}{2}-\pi\sum_{k=1}^{n}(a_k^2+b_k^2)$$

(cf. 2, **147**, formulae 33 and 36). On the other hand, from Parseval's equation (cf. 2, **147** formulae 40 and **155**)

$$\int_0^{2\pi}f^2(x)dx=\frac{\pi a_0^2}{2}+\pi\sum_{k=1}^{\infty}(a_k^2+b_k^2).$$

Comparison of the last two equations shows that $\|f-S_n\|\to 0$ i.e., that $S_n(x)\to f(x)$ in the mean.

In the following whenever we consider successive functions $u_n(P)$ converging in the mean to a function $u(P)$, it will be assumed that the functions $u_n(P)$ as well as the function $u(P)$ have finite norm.

We shall denote the fact that $u_n(P)$ converges in the mean to $u(P)$ as follows:

$$u_n(P)\xrightarrow{\text{mean}}u(P). \tag{2}$$

We now establish some properties of sequences which converge in the mean.

THEOREM 1. *One and the same sequence cannot converge in the mean to two different functions.*

We assume the converse and let the sequence $u_n(P)$ converge in the mean both to the function $u(P)$ and to the function $v(P)$, so that

$$\lim_{n\to\infty}\|u_n-u\|=\lim_{n\to\infty}\|u_n-v\|=0.$$

We now estimate the norm of the difference $u-v$. By the triangle inequality

$$\|u-v\|=\|u-u_n-(v-u_n)\|\leq\|u-u_n\|+\|v-u_n\|\underset{n\to\infty}{\to}0.$$

Hence it follows that $\|u-v\|=0$ and so by property (a) of the norm (§ 3) $u(P)\equiv v(P)$.

THEOREM 2. *Let* $u_n(P) \xrightarrow{\text{mean}} u(P)$ *and let* $v(P)$ *be an arbitrary function with finite norm. Then*

$$(u_n, v) \underset{n\to\infty}{\to} (u, v). \tag{3}$$

For the proof we estimate the difference

$$(u_n, v)-(u, v)=(u_n-u, v).$$

By Bunyakovsky's inequality (the generalization of Cauchy's inequality)

$$|(u_n, v)-(u, v)| \leq \|u_n-u\|\|v\|.$$

The first factor on the right tends to zero by the definition of convergence in the mean, and the second factor is constant. The theorem has thus been proved. We shall call this theorem *the theorem of the continuity of scalar product.* We note that the somewhat stronger theorem 1 § 43 usually bears this name.

Theorem 2 can be given a rather different form which is more convenient in applications. We consider the following series which is convergent in the mean

$$\sum_{n=1}^{\infty} \varphi_n(P)=u(P), \tag{4}$$

and let $u_n(P)$ be its partial sums, so that

$$u_n(P)=\sum_{k=1}^{n}\varphi_k(P); \qquad u_n(P)\xrightarrow{\text{mean}}u(P).$$

By theorem 2

$$(u_n, v)=\sum_{k=1}^{n}(\varphi_k, v) \underset{n\to\infty}{\to} (u, v)$$

or, what is the same thing

$$\sum_{n=1}^{\infty} (\varphi_n, v)=(u, v).$$

By recalling the definition of scalar multiplication [formula (1) § 2] it is possible to rewrite the last equation in the following form:

$$\sum_{n=1}^{\infty} \int_{\Omega} \varphi_n(P)v(P)\mathrm{d}\Omega = \int_{\Omega} u(P)v(P)\mathrm{d}\Omega. \tag{5}$$

Thus theorem 2 can be formulated:

A series which is convergent in the mean can be integrated term by term, when all its terms are multiplied by any function possessing a finite norm.

Let ω be an arbitrary domain, consisting of part of the domain Ω. The volume (area or length, if the domain Ω is two or one dimensional respectively) of the domain ω we shall denote by $|\omega|$. We choose a function $v(P)$, putting

$$v(P) = \begin{cases} \dfrac{1}{|\omega|}, & P \in \omega, \\ 0, & P \in \Omega - \omega. \end{cases}$$

The function $v(P)$ possesses a finite norm:

$$\|v(P)\| = \left\{ \int_{\Omega} v^2(P)\mathrm{d}\Omega \right\}^{\frac{1}{2}} = \frac{1}{\sqrt{|\omega|}}.$$

We substitute this in (3). We have

$$(u_n, v) = \int_{\Omega} u_n(P)v(P)\mathrm{d}\Omega = \int_{\omega} u_n(P)v(P)\mathrm{d}\Omega + \int_{\Omega-\omega} u_n(P)v(P)\mathrm{d}\Omega.$$

The last integral vanishes, since $v(P) = 0$ in $\Omega - \omega$. Hence

$$(u_n, v) = \int_{\omega} u_n(P)v(P)\mathrm{d}\Omega = \frac{1}{|\omega|} \int_{\omega} u_n(P)\mathrm{d}\Omega.$$

Similarly

$$(u, v) = \frac{1}{|\omega|} \int_{\omega} u(P)\mathrm{d}\Omega.$$

The expression

$$\frac{1}{|\omega|} \int_{\omega} \psi(P)\mathrm{d}\Omega$$

is called the mean value of the function $\psi(P)$ over the

domain ω. With the function $v(P)$ used here theorem 2 reduces to the following:

If some sequence converges in the mean then the mean values of the terms of the sequence converge to the corresponding mean value of the limiting function of the sequence.

This assertion plays an essential role in mathematical physics. We are accustomed to considering physical quantities such as point functions, but it is, however, impossible to measure directly the value of such quantities at a distinct point; measurement only gives the mean value of the quantity over some small region containing the given point. Thus, if we measure the temperature of the surface of a non-uniformly heated body, then such a measurement cannot determine the temperature at a point on the surface; it can only give the mean magnitude of the temperature over some portion of the surface. The same thing occurs in the tensometric measurement of deformation or the determination of stress on the surface of a body caused by the application of an external force and also in other problems. It can be asserted that measurement generally gives not the values of physical magnitudes at a point, but the mean values of these magnitudes over some small region, and that, in practice, it is sufficient to know just this mean value. As we shall see below variational methods allow us to approximate in the mean to the quantity required. From our assertion it follows that direct methods allow us to approximate to the mean (in the above sense) values of the quantity required with an error as small as we please in absolute value. This suffices for practical purposes.

It is sometimes necessary to introduce a somewhat more general type of convergence.

Let $\sigma(P)$ be some non-negative function. We shall say that $\varphi_n(P)$ converges, as $n \to \infty$, to $\varphi(P)$ *in the mean with a weight function* $\sigma(P)$, if

$$\lim_{n\to\infty} \int_{\Omega} \sigma(P) |\varphi_n(P) - \varphi(P)|^2 \, d\Omega = 0.$$

All that has been said here about convergence in the mean applies to convergence in the mean with a weight function $\sigma(P)$.

THEOREM 3. *If* $u_n(P) \xrightarrow{\text{mean}} u(P)$, *then* $\|u_n\| \to \|u\|$.

We have $\|u_n - u\| \to 0$. By the triangle inequality

$$| \|u_n\| - \|u\| | \leq \|u_n - u\|.$$

Hence $| \|u_n\| - \|u\| | \to 0$ and consequently, $\|u_n\| \to \|u\|$.

It is possible to consider functions defined not in a region, but on a surface or a curve, and for such functions also to define convergence in the mean. Let, for example, the functions $u(Q)$ and $u_n(Q)$, $n = 1, 2, \ldots$, be defined on some surface S (Q is a variable point on this surface). We shall say that $u_n(Q) \xrightarrow{\text{mean}} u(Q)$ on the surface S, if all the integrals

$$\int_S u^2(Q)\mathrm{d}S; \quad \int_S u_n^2(Q)\mathrm{d}S, \quad n = 1, 2, \ldots,$$

exist and if

$$\int_S [u(Q) - u_n(Q)]^2 \mathrm{d}S \underset{n\to\infty}{\to} 0.$$

All that has been said until now about convergence in the mean is still applicable in this case.

In order to explain more clearly the relationship between convergence in the mean and uniform convergence we consider an example. The series

$$\sum_{n=1}^{\infty} \frac{\cos nx}{n^2} \tag{6}$$

converges uniformly, say, in the interval $0 \leq x \leq \pi$. In fact

$$\left| \frac{\cos nx}{n^2} \right| \leq \frac{1}{n^2},$$

and the series $\sum_{n=1}^{\infty} \frac{1}{n^2}$ converges; the uniform convergence of series (6) *follows* from Weierstrass' criterion (cf 1, **147**). *A fortiori* (see below), the series converges in the mean in the interval mentioned.

We now compare the rapidity of uniform convergence and of convergence in the mean for series (6). To this end we estimate both the absolute magnitude and the norm of the remainder

$$R_N = \sum_{n=N+1}^{\infty} \frac{\cos nx}{n^2};$$

the first characterizes the rapidity of uniform convergence, the second — the rapidity of convergence in the mean. We have

$$|R_N| \leq \sum_{n=N+1}^{\infty} \frac{1}{n^2}.$$

The equality sign is possible here and occurs at $x=0$. Using the integral test for the convergence of series (cf. 1, **122**), we readily find that

$$\sum_{n=N+1}^{\infty} \frac{1}{n^2} > \int_{N+1}^{\infty} \frac{dn}{n^2} = \frac{1}{N+1},$$

and consequently, at least for $x=0$

$$|R_N| > \frac{1}{N+1}. \tag{7}$$

Further

$$\|R_N\|^2 = \int_0^{\pi} \left[\sum_{n=N+1}^{\infty} \frac{\cos nx}{n^2} \right]^2 dx,$$

or, since $\cos nx$ is an even function of x

$$\|R_N\|^2 = \frac{1}{2} \int_{-\pi}^{\pi} \left[\sum_{n=N+1}^{\infty} \frac{\cos nx}{n^2} \right]^2 dx.$$

By the Parseval formula, (cf., 2, **147** and **155**) we find

$$\|R_N\|^2 = \frac{\pi}{2} \sum_{n=N+1}^{\infty} \frac{1}{n^4}.$$

Again using the integral test for convergence of series, we obtain

$$\|R_N\|^2 < \frac{\pi}{2} \int_N^{\infty} \frac{dn}{n^4} = \frac{\pi}{6N^3} < \frac{1}{N^3}$$

and finally

$$\|R_N\| < \frac{1}{N^{3/2}}. \tag{8}$$

Formulae (7) and (8) prove that convergence in the mean is faster than uniform convergence. This conclusion, which we have established by a particular example, is in fact of general character. In particular *convergence in the mean always follows from uniform convergence.* In fact, let $u_n(x) \to u(x)$ uniformly in the interval $a \leq x \leq b$. This means that for a given $\varepsilon > 0$ it is possible to find an $N(\varepsilon)$ such that if $n \geq N$

$$|u_n(x) - u(x)| < \frac{\varepsilon}{\sqrt{(b-a)}}.$$

Squaring and integrating we find that

$$\int_a^b [u_n(x)-u(x)]^2 dx < \varepsilon^2,$$

or $\|u_n-u\|<\varepsilon$, $n\geq N(\varepsilon)$ and this means that $u_n \xrightarrow{\text{mean}} u$. The converse conclusion is invalid: a series which converges in the mean need not converge uniformly. Thus the Fourier series for the function $\ln 2\left|\sin\frac{x}{2}\right|$ has the form

$$\ln 2\left|\sin\frac{x}{2}\right| = -\sum_{n=1}^{\infty}\frac{\cos nx}{n}; \tag{9}$$

it converges in the mean, e. g. in the segment $-\pi\leq x\leq\pi$, since the function $\ln 2\left|\sin\frac{x}{2}\right|$ has a finite norm. At the same time series (9) does not converge uniformly in the interval indicated. Moreover, this series diverges in fact for $x=0$.

§ 8. Energy convergence

In the present paragraph we shall study another type of convergence connected with a new notion of the measure of the closeness of two functions. The new type of convergence plays an important role in the theory of variational methods.

We consider a *positive definite* operator A and its field of definition D_A. If the operator A is differential, as it will most frequently be in the following, then we shall consider that functions within D_A are continuous and have the necessary number of continuous derivatives in some closed domain $\bar{\Omega}=\Omega+S$. More precisely, if k is the order of the operator A, i. e. the order of the highest derivative involved in A, then we shall assume that the functions themselves from D_A and their derivatives up to order $k-1$ inclusive are continuous in $\bar{\Omega}$ and the derivatives of order k are continuous in the open domain Ω, with the exception, possibly, of a finite number of points, curves or surfaces, and that they have a finite norm. It is easily seen that Au then also has a finite norm if $u\in D_A$. Besides the continuity conditions mentioned the functions

from the field of definition of the differential operator also satisfy the boundary conditions of the corresponding problem.

Concerning these conditions we make the following assumption: we shall consider that the boundary conditions are homogeneous, i. e. such that any function which is identically equal to zero both on the boundary S itself and at all the points of the domain Ω sufficiently close to S, satisfies them.

In the following, when considering the field of definition of a differential operator, we shall not make special mention of the conditions of continuity and differentiability, but will take them as understood, formulating only the boundary conditions.

Let $u(P)$ and $v(P)$ be two functions from D_A. *We take as a measure of the closeness of these functions the square root of the energy of their difference.* For brevity we denote the square root of the energy of some function $u(P) \in D_A$ by $\|u\|$, so that

$$\|u\| = \sqrt{(Au, u)}. \tag{1}$$

As has been said, we adopt the quantity

$$\|u-v\| = \sqrt{(A(u-v), (u-v))}$$

as a measure of the closeness of, or distance between, the functions $u(P)$ and $v(P)$.

We shall call the quantity $\|u\|$ the *norm of the energy* of the function $u(P)$. This term will be explained later.

If it becomes necessary to stress the role of the operator A then we shall write $\|u\|_A$ in place of $\|u\|$.

A new formulation of the distance between functions leads, as was mentioned in § 7, to a new concept of the convergence of functions. Let $u(P) \in D_A$ and $u_n(P) \in D_A$, $n = 1, 2, \ldots$. We shall say that *the sequence $u_n(P)$ converges in energy to $u(P)$, if* $\|u_n - u\| \xrightarrow[n\to\infty]{} 0$.

The fact that $u_n(P)$ converges in energy to $u(P)$ we shall denote symbolically thus

$$u_n(P) \overset{E}{\to} u(P).$$

It must be stressed that the concepts of energy, energy convergence etc., are defined only after the corresponding positive definite operator A has been specified.

From the definition there immediately follows:

THEOREM 1. *If the given operator is positive-bounded-below and* $u_n \xrightarrow{E} u$, *then it follows that* $u_n \xrightarrow{\text{mean}} u$.

In fact, if A is a positive-bounded-below operator, then for any function $u(P) \in D_A$ the inequality $(Au, u) \geq \gamma^2 \|u\|^2$ is valid; that is, in the new notation

$$|u| \geq \gamma \|u\|, \qquad u \in D_A. \tag{2}$$

Replacing u by $u_n - u$, we obtain

$$\|u_n - u\| \leq \frac{1}{\gamma} |u_n - u| \underset{n \to \infty}{\to} 0.$$

Convergence in the mean, as studied in the previous section, is a particular case of energy convergence. If the operator A is the identity operator E, $(Au, u) = (u, u) = \|u\|^2$ and so $|u| = \|u\|$ and the condition of energy convergence $|u_n - u| \to 0$ is identical with the condition $\|u_n - u\| \to 0$. Another particular case of energy convergence is convergence in the mean with a weighting function, provided that its weighting function $\sigma(P)$ does not vanish everywhere and that it does not vanish at more than a finite number of points, or on a finite number of curves or surfaces. In order to establish this we take as the operator A, multiplication by the function $\sigma(P)$:

$$Au = \sigma(P) u(P).$$

We now consider some less simple examples. Let B be the operator introduced in § 5.

$$Bu = -\frac{d^2u}{dx^2}, \quad 0 < x < 1; \quad u(0) = u(1) = 0.$$

For this operator, as we have seen,

$$|u|^2 = (Bu, u) = \int_0^1 u_x^2(x)\, dx.$$

The energy convergence, $u_n \xrightarrow{E} u$, in the given case means that

$$\int_0^1 [u_{An}(x) - u_A(x)]^2 dx \underset{n\to\infty}{\to} 0, \tag{3}$$

i.e. that the first derivative of $u_n(x)$ converges in the mean to the first derivative of the function $u(x)$. In § 6 it was explained that B is a positive-bounded-below operator; from theorem 1 it follows that if $u_n(x) \xrightarrow{E} u(x)$ then simultaneously $u_n(x) \xrightarrow{\text{mean}} u(x)$. It is easily shown however, that in the example considered $u_n(x) \to u(x)$ *uniformly* in the interval $0 \leq x \leq 1$. For the proof, we note, first of all, that in virtue of the boundary conditions, $u_n(0) - u(0) = 0$. Thus

$$u_n(x) - u(x) = \int_0^x [u_{An}(t) - u_A(t)] dt.$$

We estimate the integral by Bunyakovsky's inequality, taking one of the integrands on the right-hand side as unity. This gives

$$|u_n(x) - u(x)| \leq \left\{\int_0^x 1^2 dt\right\}^{\frac{1}{2}} \left\{\int_0^x [u_{An}(t) - u_A(t)]^2 dt\right\}^{\frac{1}{2}}$$

$$= \sqrt{x}\left\{\int_0^x [u_{An}(t) - u_A(t)]^2 dt\right\}^{\frac{1}{2}} \leq \left\{\int_0^1 [u_{An}(t) - u_A(t)]^2 dt\right\}^{\frac{1}{2}}.$$

The latter expression does not depend upon x and tends to zero in virtue of condition (3). Thus $u_n(x) \to u(x)$ uniformly, by the well-known Weierstrass test. (For the Weierstrass test for series, cf. 1, 147. The reader will not find it difficult to formulate and prove the analogous test for sequences.)

As the second example we consider the operator $-\nabla^2$ (∇^2 is the Laplacian) operating on functions which vanish at the boundary S of some domain Ω. We have seen (§ 6) that this operator is positive definite. Later (see § 22) we shall prove

that it is also positive-bounded-below. The energy of a function $u(P) \in D$ is (see § 6)

$$\|u\|^2 = (-\nabla^2 u, u) = \int_\Omega \sum_{i=1}^{m} \left(\frac{\partial u}{\partial x_i}\right)^2 d\Omega.$$

If $u_n(P) \xrightarrow{E} u(P)$, then for this operator this means that

$$\int_\Omega \sum_{i=1}^{m} \left(\frac{\partial u_n}{\partial x_i} - \frac{\partial u}{\partial x_i}\right)^2 d\Omega \underset{n\to\infty}{\to} 0.$$

Hence it easily follows that

$$\int_\Omega \left(\frac{\partial u_n}{\partial x_i} - \frac{\partial u}{\partial x_i}\right)^2 d\Omega \underset{n\to\infty}{\to} 0, \quad i = 1, 2, \ldots, m.$$

Thus in the given case the energy convergence means that the first derivatives of the function $u_n(P)$ tend in the mean to the corresponding first derivatives of the function $u(P)$. From the fact that the operator ∇^2 is positive-bounded-below it follows, in virtue of theorem 1, that $u_n(P)$ converges in the mean to $u(P)$ over the domain Ω.

The following stronger assertion is also valid: in the case considered it also follows from the relationship $u_n(P) \xrightarrow{E} u(P)$ that $u_n(P)$ converges in the mean to $u(P)$ over any piecewise smooth surface (curve – if the domain Ω is planar) which lies entirely within Ω; in particular, u_n converges to u in the mean over the boundary of the domain Ω.

Assuming the operator A, to be positive definite let us consider the scalar product (Au, v). It is meaningful if $u \in D_A$, and v is an arbitrary function with a finite norm. We shall consider this product on the assumption that not only u but also v belongs to the field of definition of the operator A. In particular v satisfies the same boundary conditions as the function u. In this case we shall call the quantity (Au, v)

the energy product of the functions $u(P)$ and $v(P)$ and denote it by the symbol $[u, v]$ so that

$$[u, v]=(Au, v)=\int_{\Omega} vAu\,d\Omega;\ u\in D_A,\ v\in D_A. \tag{4}$$

It is remarkable that *the energy product possesses the same properties* A–D *as the scalar product of functions* (see § 3). We shall verify this. We have

$$[v, u]=(Av, u).$$

The operator A, is positive definite and symmetric, therefore $(Av, u)=(v, Au)=[u, v]$ and property A holds. Property B follows from the linearity of the operator A, and properties C and D from the positive definite property.

We further note that for the identity operator the energy product becomes the scalar product of the functions.

From formulae (1) and (4) it is clear that the energy product is connected with the norm of the energy in the same way as is the scalar product with the ordinary norm. By repeating the reasoning of § 3 we easily convince ourselves that the norm of the energy possesses properties (a)–(d) of the ordinary norm. In particular there holds the inequality

$$|[u, v]|\leq |u||v|, \tag{5}$$

analogous to the Cauchy - Bunyakovsky inequality, and also the triangle inequality

$$|u+v|\leq |u|+|v|. \tag{6}$$

§ 9. The linear independence of functions

The functions $u_1(P)$, $u_2(P)$, ..., $u_n(P)$ are called *linearly dependent* if it is possible to find constants a_1, a_2, ..., a_n, not all equal to zero, such that the following identity is satisfied

$$a_1u_1(P)+a_2u_2(P)+\ \ldots\ +a_n(P)\equiv 0. \tag{1}$$

These functions are called *linearly independent* if an identity of this form holds only if

$$a_1 = a_2 = \ldots = a_n = 0.$$

The concept of linear dependencé or independence is familiar to the reader from the theory of ordinary linear differential equations (cf 2, 26). Thc test for linear independence of functions involving the non-vanishing of their Wronskian determinant is also familiar. We now give another such test which is considerably more convenient for future applications.

THEOREM 1. *Let the functions* $u_1(P)$, $u_2(P)$, ..., $u_n(P)$ *be specified in some domain* Ω *and belong to the field of definition of some positive operator* A. *In order that these functions be linearly dependent it is necessary and sufficient that the determinant*

$$\begin{vmatrix} [u_1, u_1]; & [u_1, u_2]; & \ldots & [u_1, u_n] \\ [u_2, u_1]; & [u_2, u_2]; & \ldots & [u_2, u_n] \\ , & \ldots & \ldots & \ldots \\ [u_n, u_1]; & [u_n, u_2]; & \ldots & [u_n, u_n] \end{vmatrix} \tag{2}$$

vanish.

Determinant (2) is known as the *Gram determinant* of the functions u_1, u_2, ..., u_n. We recall that

$$[u_i, u_k] = (Au_i, u_k) = \int_\Omega u_k A u_i \mathrm{d}\Omega$$

is the energy product of the functions $u_i(P)$ and $u_k(P)$.

Proof (the necessity). Let the functions $u_1(P)$, $u_2(P)$, ..., $u_n(P)$ be linearly dependent. Then there exist constants a_1, a_2, ..., a_n, not all equal to zero, such that

$$a_1 u_1(P) + a_2 u_2(P) + \ldots + a_n u_n(P) = 0.$$

We form the energy product of both sides of this equation with Au_k where k is one of the numbers 1, 2, ..., n;

$$[u_k, u_1]a_1 + [u_k, u_2]a_2 + \ldots + [u_k, u_n]a_n = 0.$$

By giving k the values 1, 2, ..., n, we obtain the system of equations

$$\begin{aligned}
&[u_1, u_1]a_1+[u_1, u_2]a_2+ \ldots +[u_1, u_n]a_n=0,\\
&[u_2, u_1]a_1+[u_2, u_2]a_2+ \ldots +[u_2, u_n]a_n=0,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&[u_n, u_1]a_1+[u_n, u_2]a_2+ \ldots +[u_n, u_n]a_n=0.
\end{aligned} \tag{3}$$

This can be considered as a system of n homogeneous equations in the n unknowns a_1, a_2, ..., a_n. As we know, this system admits of a non-trivial solution and so, the determinant of this system, which is in fact the Gram determinant of the functions $u_1(P)$, $u_2(P)$, ..., $u_n(P)$ is zero.

Proof (the sufficiency). Let determinant (2) equal zero, and let us construct system (3).

It has a solution in which not all the unknowns a_k equal zero. Suppose that such a solution is a_1, a_2, ..., a_n. We let

$$a_1u_1+a_2u_2+ \ldots +a_nu_n=u;$$

properties A–D of scalar product apply also for the energy products, and so by property B of scalar product, equations (3) can be represented in the form

$$[u_1, u]=0; \quad [u_2, u]=0, \ \ldots, \ [u_n, u]=0.$$

We multiply the first equation by a_1 the second by a_2 etc, and add the equation so obtained. In virtue of property B, the result can be reduced to the form $[u, u]=0$, and by property D of the scalar product $u\equiv 0$, which was to be proved.

From theorem 1 it immediately follows that a necessary and sufficient condition for the linear independence of a set of functions is that their Gram determinant be different from zero.

We consider the particular case in which the operator A is the identity operator. The condition that the functions belong to the field of definition D_A of the operator then reduces to the fact that they must simply have a finite norm; the energy and scalar products coincide. As a direct consequence of theorem 1 we then have:

THEOREM 2. *In order that the functions* $u_1(P)$, $u_2(P)$, ..., $u_n(P)$ *which possess a finite norm, be linearly dependent it is necessary and sufficient that the determinant*

$$\begin{vmatrix} (u_1, u_1); (u_1, u_2); \ldots (u_1, u_n) \\ (u_2, u_1); (u_2, u_2); \ldots (u_2, u_n) \\ \ldots\ldots\ldots\ldots\ldots \\ (u_n, u_1); (u_n, u_2); \ldots (u_n, u_n) \end{vmatrix} \tag{4}$$

vanish.

In virtue of the restrictions imposed in § 5, the functions within the field of definition of some positive operator in each case have a finite norm, and so it is possible to apply both theorems of this section to such functions.

In conclusion we note that a function identically equal to zero is linearly dependent with respect to all other functions.

§ 10. Orthogonality and orthogonal series

Two functions are termed *orthogonal over the domain* Ω if their scalar product with respect to this domain vanishes. Thus the functions sin x and cos x are orthogonal over the interval $0<x<2\pi$ because

$$\int_0^{2\pi} \sin x \cos x \mathrm{d}x = 0.$$

In precisely the same way the functions x^2-y^2 and xy are orthogonal over the circle $x^2+y^2<1$ because

$$\iint_{x^2+y^2<1} (x^2-y^2)xy\mathrm{d}x\mathrm{d}y = \frac{1}{2}\int_0^1 \varrho^5 \mathrm{d}\varrho \int_0^{2\pi} \cos 2\vartheta \sin 2\vartheta \mathrm{d}\vartheta = 0;$$

where for the evaluation of the integral we have introduced polar coordinates by taking $x=\varrho \cos \vartheta$ and $y=\varrho \sin \vartheta$.

The property of orthogonality of trigonometric functions — which are used in the construction of the theory of Fourier

series — plays an important role in analysis. In the present section orthogonal systems of a more general type will be considered.

We now introduce the further concept of a normalized function which is defined to be a function whose norm is unity.

A finite or infinite sequence of functions is called *an orthogonal set* if the functions of this set are mutually orthogonal. Such a set is further *orthonormal* if all the functions of the set are normalized. By definition, if $\varphi_1(P)$, $\varphi_2(P)$, $\ldots$, $\varphi_n(P)$, $\ldots$ is an orthonormal set in the domain Ω, then

$$(\varphi_k, \varphi_m) = \begin{cases} 0, & k \neq m, \\ 1, & k = m; \end{cases} \tag{1}$$

the scalar product in (1) being taken with respect to the domain Ω.

Any orthogonal set is easily converted into an orthonormal one as all that is necessary is to divide each function of the set by its norm (which we shall always assume to be finite and different from zero).

The functions of an orthonormal set are linearly independent. In fact, if $\varphi_1, \varphi_2, \ldots, \varphi_n$ are normalized and mutually orthogonal then, in their Gram determinant, the diagonal elements equal unity and the off-diagonal elements equal zero, and so the Gram determinant becomes unity. Thus, by theorem 2 of § 9 the functions $\varphi_1, \varphi_2, \ldots, \varphi_n$ are linearly independent.

If $\varphi = \varphi_1 + \varphi_2 + \ldots + \varphi_n$, and the terms are mutually orthogonal, then

$$\|\varphi\|^2 = \|\varphi_1\|^2 + \|\varphi_2\|^2 + \ldots + \|\varphi_n\|^2. \tag{2}$$

The proof of this equality, which is an important generalization of Pythagoras's theorem, is very simple

$$\|\varphi\|^2 = (\varphi, \varphi) = \left(\sum_{k=1}^{n} \varphi_k, \sum_{m=1}^{n} \varphi_m \right) = \sum_{k,m=1}^{n} (\varphi_k, \varphi_m).$$

If $k \neq m$ then $(\varphi_k, \varphi_m) = 0$, and the double summation reduces to a simple summation so that

$$\|\varphi\|^2 = \sum_{k=1}^{n} (\varphi_k, \varphi_k),$$

which, by virtue of the definition of a norm, gives

$$\|\varphi\|^2 = \sum_{k=1}^{n} \|\varphi_k\|^2.$$

Formula (2) is easily extended to the case when φ is the sum of an infinite series of orthogonal functions, provided that this series converges in the mean. Let

$$\varphi = \sum_{k=1}^{\infty} \varphi_k,$$

where the series converges in the mean, and let the terms of this series be mutually orthogonal. We take

$$s_n = \sum_{k=1}^{n} \varphi_k,$$

so that $s_n \xrightarrow{\text{mean}} \varphi$. From what has been shown already,

$$\|s_n\|^2 = \sum_{k=1}^{n} \|\varphi_k\|^2;$$

on the other hand, by theorem 3 § 7, $\|s_n\|^2 \to \|\varphi\|^2$. Hence

$$\|\varphi\|^2 = \sum_{k=1}^{\infty} \|\varphi_k\|^2. \tag{2$_1$}$$

In particular, if the set of functions φ_n, $n = 1, 2, \ldots$ is orthonormal, and

$$\varphi = \sum_{n=1}^{\infty} a_n \varphi_n \tag{3}$$

the series being assumed to converge in the mean, then

$$\|\varphi\| = \sum_{n=1}^{\infty} a_n^2. \tag{4}$$

In passing we have obtained the result that if the series

$$\sum_{n=1}^{\infty} a_n \varphi_n,$$

where the functions $\varphi_n(P)$ are orthonormal, converges in the mean, then the series of the squares of its coefficients necessarily converges. It can be proved that this condition is also sufficient, i.e., if series (4) converges, then series (3) converges in the mean to some function the norm of which equals the sum of series (4).

Examples. 1. A classical example is the set of functions:

$$1, \cos x, \sin x, \cos 2x, \sin 2x, \ldots, \tag{5}$$

which are orthogonal in the interval $0<x<2\pi$; in virtue of the periodicity of the functions of the system they are orthogonal in any interval of length 2π. To orthonormalize system (5) it suffices to divide unity by $\sqrt{(2\pi)}$ and the remaining functions of the system by $\sqrt{(\pi)}$.

2. The set of functions

$$1, \cos x, \cos 2x, \ldots, \cos nx, \ldots \tag{6}$$

is orthogonal in the interval $0<x<\pi$. The corresponding orthonormalized set will be

$$\frac{1}{\sqrt{(\pi)}}, \sqrt{\left(\frac{2}{\pi}\right)} \cos x, \sqrt{\left(\frac{2}{\pi}\right)} \cos 2x, \ldots, \sqrt{\left(\frac{2}{\pi}\right)} \cos nx, \ldots. \tag{6_1}$$

3. In the same interval the set

$$\sin nx, \ n=1, 2, \ldots \tag{7}$$

is orthogonal. Its corresponding orthonormal set is

$$\sqrt{\left(\frac{2}{\pi}\right)} \sin nx, \ n=1, 2, \ldots. \tag{7_1}$$

4. In the square $0<x<\pi$, $0<y<\pi$ the system

$$\frac{2}{\pi} \sin mx \sin ny; \ m=1, 2, \ldots; \ n=1, 2, \ldots \tag{8}$$

for example, is orthonormal.

5. The Legendre polynomials

$$P_n(x)=\frac{1}{2^n n!} \frac{d^n(x^2-1)^n}{dx^n}, \quad n=0, 1, 2, \ldots \tag{9}$$

are orthogonal in the interval $-1<x<1$. It is well known that

$$P_n(x)\|^2=\int_{-1}^{1}P_n^2(x)\mathrm{d}x=\frac{2}{2n+1},$$

and that therefore the polynomials

$$\sqrt{\left(\frac{2n+1}{2}\right)}P_n(x),\quad n=0,\ 1,\ 2,$$

constitute the corresponding orthonormal set.

Let there be given in some domain Ω an orthonormal set of functions $\varphi_k(P)$, $k=1,\ 2,\ \ldots$, and the function $u(P)$, having a finite norm. We specify a positive integer u and pose the problem of so selecting the constants $\alpha_1, \alpha_2,\ \ldots\ \alpha_n$, that the norm $\|u-s_n\|$, where $s_n(P)=\alpha_1\varphi_1(P)+\alpha_2\varphi_2(P)+\ldots+\alpha_n\varphi_n(P)$, be a minimum.

We have

$$\begin{aligned}\|u-s_n\|^2&=(u-s_n,\ u-s_n)=(u,\ u)-2(u,\ s_n)+(s_n,\ s_n)\\&=\|u\|^2-2(u,\ s_n)+\|s_n\|^2.\end{aligned}$$

By formulae (3) and (4)

$$\|s_n\|^2=\sum_{k=1}^{n}\alpha_k^2.$$

Further

$$(u,\ s_n)=\left(u,\sum_{k=1}^{n}\alpha_k\varphi_k\right)=\sum_{k=1}^{n}\alpha_k(u,\ \varphi_k).$$

The scalar products

$$(u,\ \varphi_k)=\int_{\Omega}u(P)\varphi_k(P)\mathrm{d}\Omega \tag{10}$$

are known as *the Fourier coefficients* of the function $u(P)$ with respect to the orthonormal set of functions $\varphi_n(P)$. For brevity we use the notation

$$(u,\ \varphi_k)=a_k,$$

then

$$(u,\ s_n)=\sum_{k=1}^{n}\alpha_k a_k$$

and

$$\|u-s_n\|^2=\|u\|^2+\sum_{k=1}^{n}(\alpha_k^2-2\alpha_k a_k).$$

By adding and subtracting the quantity a_k^2 under the sign of summation we obtain

$$\|u-s_n\|^2=\|u\|^2+\sum_{k=1}^{n}(\alpha_k-a_k)^2-\sum_{k=1}^{n}a_k^2.$$

From the latter equality it is clear that the quantity $\|u-s_n\|$ assumes its minimum value if we take $\alpha_k=a_k$, i.e. if we take as the constants α_k the Fourier coefficients of the function $u(P)$. That is

$$s_n(P)=\sum_{k=1}^{n}a_k\varphi_k(P)=\sum_{k=1}^{n}(u,\varphi_k)\varphi_k(P) \tag{11}$$

and

$$\|u-s_n\|^2=\|u\|^2-\sum_{k=1}^{n}a_k^2. \tag{12}$$

The left-hand side of equation (12) is non-negative. There thus follows the important inequality

$$\sum_{k=1}^{n}a_k^2\leq\|u\|^2. \tag{13}$$

Since n can be chosen arbitrarily, we let $n\to\infty$. From inequality (13) it follows that the series

$$\sum_{k=1}^{\infty}a_k^2 \tag{14}$$

converges, and that

$$\sum_{k=1}^{\infty}a_k^2\leq\|u\|^2. \tag{15}$$

Inequality (15) is called *Bessel's inequality.*

The formal transition to $n=\infty$ in the sum s_n leads to the series

$$\sum_{k=1}^{\infty}a_k\varphi_k(P)=\sum_{k=1}^{\infty}(u,\varphi_k)\varphi_k(P). \tag{16}$$

This is known as the *Fourier series* or *orthogonal series* of the function $u(P)$ with respect to the orthonormal system of functions $\varphi_n(P)$. It is of importance to consider the conditions under which series (16) converges to the function $u(P)$.

We shall introduce the concept of the *complete set of functions*. The set of functions $\varphi_n(P)$, $n=1, 2, \ldots$ (whether orthogonal or not is immaterial) is said to be complete in the domain Ω in the sense of convergence in the mean, if any function with a finite norm, can be, to within an arbitrary degree of precision, approximated to in the mean within Ω by linear combinations of a finite number of functions $\varphi_n(P)$. Thus according to the definition given, if the system of functions $\varphi_n(P)$ is complete, and $u(P)$ is an arbitrary function with a finite norm then for any $\varepsilon>0$ it is possible to find a natural number N and constants $\alpha_1, \alpha_2, \ldots, \alpha n$ such that

$$\|u-(\alpha_1\varphi_1+\alpha_2\varphi_2+\ldots+\alpha_N\varphi_N)\|=\left\|u-\sum_{k=1}^{N}\alpha_k\varphi_k\right\|<\varepsilon. \qquad (17)$$

It is possible to restrict the definition of completeness somewhat and to speak of *a set that is complete with respect to a given class of functions* if the inequality (17) can be constructed for any function $u(P)$ of the given class.

Thus it is easy to prove that the powers of x

$$1, x, x^2, \ldots, x^n, \ldots$$

form a complete set with respect to the class of functions which are continuous in a specified interval $a\leq x\leq b$. In fact, let $u(x)$ be such a function. Linear combinations of the powers are simply polynomials. By the well known Weierstrass theorem (cf. 2, **154**) it is possible to construct a sequence of polynomials which converges uniformly to $u(x)$ for $a\leq x\leq b$.

Hence it follows that for a specified $\varepsilon>0$ it is possible to find a polynomial $P(x)$ such that

$$|u(x)-P(x)|<\frac{\varepsilon}{\sqrt{(b-a)}}, \quad a\leq x\leq b.$$

Squaring and integrating in the range from a to b, we obtain

$$\int_a^b (u(x) - P(x))^2 dx < \varepsilon^2,$$

or $\|u - P\| < \varepsilon$, which was to be proved. Similarly it is possible to prove that the set (6) is complete when considering the class of functions which are continuous and periodic with period 2π. It is considerably more difficult, although it is possible, to prove the completeness of the set mentioned when considering the class of function with a finite norm.

THEOREM 1. *If an orthonormal set of functions $\varphi_n(P)$ is complete, in the sense of convergence in the mean, with respect to some class of functions, then the Fourier series of any function $u(P)$ of the given class*

$$\sum_{n=1}^{\infty} a_n \varphi_n(P) = \sum_{n=1}^{\infty} (u, \varphi_n)\varphi_n(P)$$

converges in the mean to this function. In this case there occurs the so called Parseval equation.

$$\|u\|^2 = \sum_{n=1}^{\infty} a_n^2 = \sum_{n=1}^{\infty} (u, \varphi_n)^2. \tag{18}$$

In virtue of the completeness of the system it is possible for a specified ε to find the numbers N, $\alpha_1, \alpha_2, \ldots, \alpha_N$ such that inequality (17) is satisfied. The quantity $\|u - \sum_{k=1}^{N} \alpha_k \varphi_k\|$ is minimized for $\alpha_k = a_k = (u, \varphi_k)$. *A fortiori*,

$$\left\| u - \sum_{k=1}^{N} a_k \varphi_k \right\| < \varepsilon.$$

From formula (12) it follows that the quantity $\|u - s_n\|$ where $s_n = \sum_{k=1}^{N} a_k \varphi_k$, decreases (at least, does not increase) with increasing n so that

$$\left\| u - \sum_{k=1}^{n} a_k \varphi_k \right\| < \varepsilon, \quad n \geq N.$$

The latter relation also means that the Fourier series converges in the mean to the function $u(P)$. We obtain the Parseval equation by letting $n \to \infty$ in (12) and recalling that, because it has been proved that the Fourier series converges, $\|u - s_n\| \to 0$.

It will be readily noticed that, in the reasoning of the present paragraph, we have only made use of the properties A–D of the scalar product and properties (a) – (d) of the norm. The fact that the scalar product is the integral of the product of two functions has played no part whatsoever. We can, therefore, generalize the concept of orthogonality and the theory of orthogonal series by taking as a foundation not the scalar but the energy product of two functions. As we shall see below, this generalization will prove extremely valuable to us.

We shall consider only functions which enter into the field of definition D_A of some positive definite operator A; we introduce the corresponding energy product $[u, v] = (Au, v)$ and the energy norm $|u| = \sqrt{(Au, v)}$. We shall call two functions *orthogonal in energy* if their energy product equals zero; a function whose energy norm is unity we shall call *normalized in energy*. "Energy orthogonal" or "energy orthonormal" systems of functions are defined analogously. If the system of functions $\varphi_n(P)$ is energy orthonormal, then

$$[\varphi, \varphi_m] = \begin{cases} 0, & k \neq m, \\ 1, & k = m. \end{cases} \tag{19}$$

If $\varphi = \varphi_1 + \varphi_2 + \ldots + \varphi_n$, where the terms are mutually energy orthonormal then $|\varphi|^2 = |\varphi_1|^2 + |\varphi_2|^2 + \ldots + |\varphi_n|^2$; if

$$\varphi = \sum_{n=1}^{\infty} a_n \varphi_n \tag{20}$$

and the system φ_n is energy orthonormal, then

$$|\varphi|^2 = \sum_{n=1}^{\infty} a_n^2. \tag{21}$$

Thus, if the functions φ_n are energy orthonormal and the series (20) converges in energy, then series (21) necessarily converges.

If an energy orthonormal system of functions $\varphi_n(P)$ and some function $u(P)\in D_A$ are given, then it is possible to pose the problem of finding constants $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that the energy norm $|u-s_n|$, where $s_n(P)=\alpha_1\varphi_2(P)+\alpha_2\varphi_2(P)+ +\ldots+\alpha_n\varphi_n(P)$, be a minimum.
As before we find that

$$\alpha_k=a_k=[u, \varphi_k],$$

and in virtue of Bessel's inequality

$$\sum_{k=1}^{\infty} a_k^2=\sum_{k=1}^{\infty}[u, \varphi_k]^2\leq |u|^2. \tag{22}$$

We shall use Fourier coefficient for the quantities $[u, \varphi_k]$ also. The series

$$\sum_{k=1}^{\infty}[u, \varphi_k]\varphi_k \tag{23}$$

we shall call the energy orthogonal series of the function $u(P)$, and also the Fourier series of this function. In a similar sense to the above, we have the concept of a set of functions which is complete in energy or complete in the sense of energy convergence. Let there be given some class of functions from D_A; in particular this class can coincide with D_A. We shall call the set of functions $\varphi_n(P)\in D_A$ *complete in energy with respect to the given class of functions*, if for any $\varepsilon>0$ and any function $u(P)$ of the given class, it is possible to find a positive integer N and constants $\alpha_1, \alpha_2, \ldots, \alpha_N$ such that the following inequality is satisfied

$$|u-(\alpha_1\varphi_1+\alpha_2\varphi_2+\ \ldots\ +\alpha_N\varphi_N)|<\varepsilon. \tag{24}$$

There follows:

THEOREM 2. *If an energy orthonormal set of functions* $\varphi_n(P)$

is complete in energy with respect to some class of functions, then the Fourier series of any function $u(P)$ of this class

$$\sum_{n=1}^{\infty} [u, \varphi_n]\varphi_n(P)$$

converges in energy to this function. We have the Parseval equation:

$$|u|^2 = \sum_{n=1}^{\infty} [u, \varphi_n]^2. \tag{25}$$

We omit the proof of this theorem since it is identical word for word with the proof of theorem 1.

We note that if the operator A which defines the energy is not only positive definite but also positive-bounded-below, then the Fourier series mentioned in theorem 2 also converges in the mean.

Example. As the operator Au we take the operator $-\frac{d^2u}{dx^2}$ where the functions $u(x)$ are subject to the boundary conditions $u(0)=u(\pi)=0$. Then

$$[u, v]=-\int_0^{\pi} v\frac{d^2u}{dx^2}\,dx=-v\frac{du}{dx}\Big|_0^{\pi}+\int_0^{\pi}\frac{du}{dx}\frac{dv}{dx}\,dx=\int_0^{\pi}\frac{du}{dx}\frac{dv}{dx}\,dx$$

and

$$|u|^2=\int_0^{\pi}\left(\frac{du}{dx}\right)^2 dx.$$

It is clear thus, that in our example, the energy orthogonality of two functions means the ordinary orthogonality of their derivatives. We further recall that, in this example, energy convergence leads to the convergence in the mean of the derivatives.

The set of functions

$$\varphi_n(x)=\frac{1}{n}\sqrt{\left(\frac{2}{\pi}\right)}\sin nx, \quad n=1, 2, \ldots$$

is energy orthonormal and complete in the class D_A. In the first place

$$[\varphi_n, \varphi_m]=\frac{2}{\pi}\int \cos nx \cos mx\,dx=\begin{cases}0, & n\neq m,\\ 1, & n=m,\end{cases}$$

so that our set is energy orthonormal. Its energy completeness is proved as follows. The set (6) is complete in the sense of convergence in the mean in

the interval $0<x<\pi$ and each function $\varphi(x)$ with a finite norm may be expressed as a Fourier series

$$\varphi(x)=b_0+\sum_{n=1}^{\infty} b_n \cos nx,$$

which converges in the mean. We put $\varphi(x)=u'(x)$ where $u(x)\in D_A$. Then $u(0)=u(\pi)=0$, and

$$b_0=\frac{1}{\pi}\int_0^{\pi}\varphi(x)\mathrm{d}x=\frac{1}{\pi}\int_0^{\pi}u'(x)\mathrm{d}x=u(\pi)-u(0)=0;$$

whence

$$u'(x)=\sum_{n=1}^{\infty} b_n \cos nx=\sum_{n=1}^{\infty} b_n^* \varphi_n'(x); \quad b_n^*=\sqrt{\left(\frac{\pi}{2}\right)}\, b_n.$$

This last equation shows that the derivative of any function from D_A can be approximated in the mean with any degree of accuracy by linear combinations of the derivatives of the $\varphi_n(x)$. This means also in this case that the set $\varphi_n(x)=\frac{1}{n}\sqrt{\left(\frac{2}{\pi}\right)}\sin nx$ is complete in energy.

Ordinary orthogonality is the particular case of energy orthogonality when A is the identity operator. Another interesting particular case is obtained if A is the operator of multiplication by the non-negative function $\sigma(P)$. In this case

$$[u, v]=\int_{\Omega}\sigma(P)u(P)v(P)\mathrm{d}\Omega;$$

the energy orthogonality of the functions $u(P)$ and $v(P)$ means that

$$\int_{\Omega}\sigma(P)u(P)v(P)\mathrm{d}\Omega=0,$$

this is the so called *orthogonality with a weight function* $\sigma(P)$. We now cite two examples of systems of functions that are orthogonal with a weight function.

1. Let $\alpha_{k,n}$; $n=1, 2, \ldots$ be the positive roots of the Bessel function $J_k(x)$. The set $J_k(\alpha_{k,n}x)$; $n=1, 2, \ldots$ is orthogonal with weight x in the interval $0<x<1$ provided that $k>-1$.

2. The Chebyshev polynomials

$$T_n(x)=\frac{1}{2^{n-1}}\cos(n\arccos x),\quad n=0,\ 1,\ 2,\ \ldots$$

are orthogonal with a weight function $\frac{1}{\sqrt{(1-x^2)}}$ in the interval $-1<x<1$.

An essential role is played in the theory of orthogonal series by the so called *process of orthonormalization*, which serves to transform any sequence of linearly independent functions into an orthonormal one. Let $\varphi_1(P), \varphi_2(P), \ldots$ be a finite or infinite sequence of functions belonging to the field of definition of some positive operator and let the functions $\varphi_1(P)$, $\varphi_2(P), \ldots, \varphi_n(P)$ be linearly independent for any n.

We construct the energy orthonormal sequence $\omega_1(P)$, $\omega_2(P), \ldots, \omega_n(P), \ldots$ such that for any n the function $\omega_n(P)$ is a linear function of $\varphi_1(P),\varphi_2(P), \ldots, \varphi_n(P)$ and conversely $\varphi_n(P)$ is a linear function of $\omega_1(P),\omega_2(P), \ldots, \omega_n(P)$.

Since the functions $\varphi_1(P), \varphi_2(P), \ldots, \varphi_n(P)$ are linearly independent $\varphi_1(P)\not\equiv 0$ and thus $|\varphi_1|>0$. We put

$$\psi_1(P)=\varphi_1(P),\quad \omega_1(P)=\frac{\psi_1(P)}{|\psi_1|}.$$

Thus the function $\omega_1(P)$ is normalized. We now suppose that the orthonormal functions $\omega_1(P), \omega_2(P), \ldots, \omega_{n-1}(P)$ have been constructed. We put

$$\psi_n(P)=\varphi_n(P)-\sum_{k=1}^{n-1}[\varphi_n, \omega_k]\omega_k(P).$$

Obviously the function $\psi_n(P)$ is orthogonal to $\omega_1(P), \omega_2(P), \ldots, \omega_{n-1}(P)$. Further, $\psi_n(P)\not\equiv 0$ otherwise the functions $\varphi_1(P), \varphi_2(P), \ldots, \varphi_n(P)$ would be linearly dependent. Now it is sufficient to put

$$\omega_n(P)=\frac{\psi_n(P)}{|\psi_n|}.$$

By construction the sequence $\omega_1(P), \omega_2(P), \ldots$ is energy orthonormal and $\omega_n(P)$ is a linear function of $\varphi_1(P), \varphi_2(P), \ldots, \varphi_n(P)$. In its turn $\varphi_n(P)$ is a linear function of $\omega_1(P), \omega_2(P_1), \ldots, \omega_n(P)$:

$$\varphi_n(P) = \| \psi_n \| \omega_n(P) + \sum_{k=1}^{n-1} [\varphi_n, \omega_k] \omega_k(P).$$

We have explained the process of orthogonalization applied to energy orthogonality. This process can be applied very simply to ordinary orthogonality, it being sufficient to assume that the positive definite operator defining the energy is the identity operator.

In concluding this section we consider some properties of orthogonal systems. For exactitude we shall speak of orthogonality in the ordinary sense, although the whole is true without change for energy orthogonality.

We shall call an orthogonal set *closed* with respect to some class of functions if for any function of this class Parseval's equation (18) holds. From theorem 1 it follows that the completeness in some class of function necessitates its being closed with respect to this class. The converse is also true: a system which is closed with respect to some class of functions is complete with respect to the same class. In fact, for any function $u(P)$ from the class mentioned, the identity (18) is valid. We denote the nth partial sum of the orthogonal series of the function by s_n. By formula (12)

$$\|u - s_n\|^2 = \|u\|^2 - \sum_{k=1}^{n} a_k^2.$$

In virtue of the Parseval equation $\|u - s_n\| \xrightarrow[n \to \infty]{} 0$. This means that the function s_n, being a linear combination of the functions $\varphi_1, \varphi_2, \ldots, \varphi_n$, for sufficiently large n approximates in the mean to the function u to within any arbitrary degree of accuracy. Since u is any function of the given class, the set of functions φ_n is complete with respect to the given class.

If an orthonormal system is complete with respect to a

given class of functions, then there does not exist in the class, a non zero function which is orthogonal to all the functions of the system. In fact if u_0 is such a function, then $(u_0, \varphi_n) = 0$, $n = 1, 2, \ldots$; by the Parseval equation

$$\|u_0\|^2 = \sum_{n=1}^{\infty} (u_0, \varphi_n)^2 = 0,$$

and by property (a) of the norm (§ 3), $u_0 \equiv 0$.

CHAPTER III

THE ENERGY METHOD

§ 11. The minimum functional theorem

Our basic problem is the solution of differential equations under associated boundary conditions. We shall consider that the boundary conditions are homogeneous; the corresponding differential equation will then, generally speaking, be non-homogeneous. We shall consider the left-hand side of the differential equation in terms of an operator defined for the set of functions which satisfy the boundary conditions. Denoting this operator by the letter A, we find that our problem consists of solving the equation

$$Au=f(P), \tag{1}$$

where $f(P)$ is a known function: *in the following we shall assume always that this function has a finite norm.* We reiterate: solving equation (1) means finding a function $u(P)\in D_A$ which turns equation (1) into an identity. The fact that the function sought belongs to the field of definition D_A of the operator A implies explicitly, that this function satisfies the boundary conditions of the given problem. Thus solving equation (1) means finding a function which satisfies both the given equation (which is most frequently a differential equation) and the boundary conditions of the problem. We shall assume that the operator A of equation (1) is positive definite.

THEOREM 1. *If the operator A is positive definite then the equation $Au=f$ cannot have more than one solution.*

Suppose that the equation mentioned has two solutions, u_1 and u_2 such that $Au_1=f$, $Au_2=f$. We put $u_1-u_2=\tilde{u}$. Subtracting and using the linearity of the operator A, we obtain $A\tilde{u}=0$. Multiplying this scalarly by $\tilde{u}$, we have $(A\tilde{u}, \tilde{u})=0$. The operator A is however positive definite and so $\tilde{u}\equiv 0$ and consequently, $u_1\equiv u_2$.

The theorems on the uniqueness of solutions of the fundamental boundary problems involving elliptical partial differential equations given in courses of mathematical physics are usually particular cases of the theorem just proved and their proof follows essentially the same lines.

The same remark applies also to Kirchhoff's theorem of the uniqueness of the solutions of the fundamental problems of static elasticity theory which is well known. See, e. g. L. S. Leibenzon [2].

THEOREM 2. *Let A be a positive definite operator, and let equation* (1) *have a solution. Then of all the values, which are given to the quadratic functional*

$$F(u)=(Au, u)-2(u, f)=\int_{\Omega}[u(P)Au-2u(P)f(P)]\mathrm{d}\Omega \qquad (2)$$

by all possible functions from the field of definition D_A of the operator A, the least is the value given to this functional by the solution of equation (1). *Conversely, if there exists in D_A a function which gives a minimal value to the functional* (2), *then this function is the solution of equation* (1).

We shall call theorem 2 *the minimal functional theorem.*

Let the function u_0 be the solution of the equation (which is unique by virtue of theorem 1) so that $Au_0=f$. Replacing f by Au_0 in (2):

$$F(u)=(Au, u)-2(Au_0, u).$$

Recalling the definition of energy product,

$$(Au, u)=[u, u], \quad (Au_0, u)=[u_0, u]$$

and

$$F(u)=[u, u]-2[u_0, u].$$

By adding and subtracting the quantity $[u_0, u_0]$, we obtain

$$F(u)=[u-u_0, u-u_0]-[u_0, u_0],$$

or

$$F(u)=\|u-u_0\|^2-\|u_0\|^2. \tag{3}$$

Hence it is clear that $F(u)$ assumes its least value when and only when $u=u_0$. Thus the first part of theorem 2 has been proved. We note that

$$\min F(u)=F(u_0)=-\|u_0\|^2=-(Au_0, u_0). \tag{4}$$

Now suppose that there exists a function $u_0(P)\in D_A$ giving a minimal value to functional (2). Let $\eta(P)$ be an arbitrary function from D_A and let λ be an arbitrary real number. Then

$$F(u_0+\lambda\eta)-F(u_0)\geq 0.$$

Using the symmetry of the operator A it is easy to reduce the last inequality to the form

$$2\lambda(Au_0-f, \eta)+\lambda^2(A\eta, \eta)\geq 0.$$

The left hand of this inequality is a quadratic in λ and is non-negative. Thus its discriminant ≤ 0 and so

$$(Au_0-f, \eta)^2-(A\eta, \eta)\cdot 0\leq 0.$$

This is only possible when

$$(Au_0-f, \eta)=0. \tag{5}$$

Putting for brevity $Au_0-f(P)=\psi(P)$, we may rewrite equation (5) in the form

$$\int_\Omega \psi(P)\eta(P)\mathrm{d}\Omega=0. \tag{5$_1$}$$

We must now prove that $\psi(P)\equiv 0$. We shall assume for the purpose of the proof that the function $\psi(P)$ is continuous. We imagine that at some point P_0 lying within Ω the function $\psi(P)$ is different from zero; say $\psi(P_0)>0$. By virtue of the continuity of the function $\psi(P)$ we can construct a small

sphere ω with centre P_0 which lies entirely within Ω (Fig. 5) and in which $\psi(P) > 0$. Both the functions which belong to D_A and their derivatives up to some order (we shall denote it by k) must be continuous, and must satisfy certain homogeneous boundary conditions on the boundary S of the

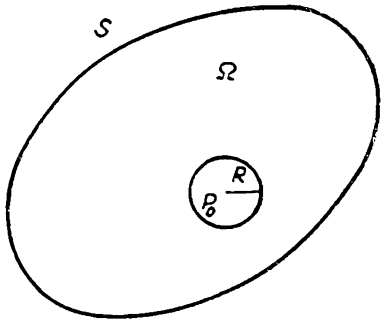

FIG. 5

domain Ω. The function $\eta(P)$ must also satisfy the same conditions but is otherwise arbitrary. With this in mind we take $\eta(P)$ as follows

$$\eta(P) = \begin{cases} (R^2 - r^2)^{k+1}, & \text{if } P \in \omega, \\ 0, & \text{if } P \in \Omega - \omega; \end{cases}$$

where R denotes the radius of the sphere ω and r the distance between the points P and P_0. It is easily seen that $\eta(P)$ is continuous in $\bar{\Omega}$ together with its derivatives up to order k inclusive; it satisfies the boundary conditions, since they are homogeneous, and the function $\eta(P) \equiv 0$ on and near S (see § 8). From what has been said it follows that $\eta(P) \in D_A$ and thus integral (5_1) must vanish. Since $\eta(P) \equiv 0$ in $\Omega - \omega$, it is sufficient to integrate over ω, and equation (5) assumes the form

$$\int_{\omega} \psi(P)(R^2 - r^2)^{k+1} \mathrm{d}\Omega = 0.$$

This is impossible because the integrand is non-negative and, at the same time, is not identically zero. Thus, it follows that $\psi(P) \equiv 0$ and consequently u_0 satisfies equation (1).

As we shall see later in a series of problems of mathematical physics the magnitude of the functional (2) is proportional to the potential energy of the system under consideration. In such cases *theorem 2 is equivalent to the principle of minimum potential energy*. This theorem allows us in the cases mentioned to replace the problem of integrating the differential equation under the specified boundary conditions by the problem of seeking a function which minimizes a functional of the form (2). The direct methods to which we shall refer later allow us to find an approximate solution to the latter problem by comparatively simple means.

The method of solving problems of mathematical physics, which instead of integration of a differential equation under specified boundary conditions involves the seeking of a function which minimizes a functional such as (2), we shall, following more or less established practice, call the *energy method*. We shall call this problem briefly the "problem of the minimum of functional (2)".

Examples. 1. We pose the problem of the integration of the equation

$$-\frac{d^2u}{dx^2}=2 \tag{6}$$

under the boundary conditions

$$u(0)=u(1)=0. \tag{7}$$

Our problem obviously has the solution $u_0=x(1-x)$. The operator defined by the left-hand side of equation (6) and by boundary conditions (7) is positive definite (actually positive-bounded-below) and so by virtue of theorem 2 the function $x(1-x)$ gives to the integral

$$F(u)=\int_0^1\left(-u\frac{d^2u}{dx^2}-4u\right)dx \tag{8}$$

a value lower than that given by any other function which satisfies the conditions (7). The requirement that the function $u(x)$ satisfy the boundary conditions is absolutely essential. Thus, if we put $u_1=x(2-x)$ violating in so

doing the condition $u(1)=0$, then we easily find that $F(u_1)=-\frac{4}{3}$ so that $F(u_0)=-\frac{1}{3}>F(u_1)$.

2. Equation (6) under the boundary conditions

$$u'(0)=0, \quad u'(1)+u(1)=0 \tag{9}$$

has the solution $u_2=3-x^2$. The operator $-\frac{d^2u}{dx^2}$ under boundary conditions (9) is positive definite (see § 6). By theorem 2, the function $3-x^2$ gives a minimum value to the integral (8) as compared with all other functions satisfying boundary conditions (9).

In connexion with the example just considered, we make the following observation. Integrating by parts and using boundary conditions (9) we may readily reduce (8) to the form

$$F(u)=\int_0^1 [u'^2(x)-4u]dx+u^2(1). \tag{10}$$

From what has been said earlier it is clear that the function $3-x^2$ gives the least value to expression (10) as compared with all the other members of the class of functions which satisfy the boundary conditions (9). It is remarkable that in this particular problem the last condition is superfluous. It can be asserted that the function $3-x^2$ gives to expression (10) a minimal value compared with any function which for $0\leq x\leq 1$ is continuous and has a continuous first derivative, independently of whether this function satisfies conditions (9) or not.

We now prove this. Let $u(x)$ be continuous and continuously differentiable for $0\leq x\leq 1$. We put $u(x)-u_2(x)=\eta(x)$, so that $u=u_2+\eta$. We have

$$\begin{aligned} F(u)=F(u_2+\eta) &= \int_0^1 [(u_2'+\eta')^2-4(u_2+\eta)]dx+[u_2(1)+u(1)]^2 \\ &= \left\{\int_0^1 (u_2'^2-4u_2)\,dx+u_2^2(1)\right\} \\ &+\left\{2\int_0^1 (u_2'\eta'-2\eta)dx+2u_2(1)\eta(1)\right\}+\left\{\int_0^1 \eta'^2dx+\eta^2(1)\right\}. \end{aligned} \tag{11}$$

In the right-hand side of (11) the first function in brackets equals $F(u_2)$ and the third bracket is positive. We shall prove that the second bracket equals zero. Integrating by parts

$$\int_0^1 u_2'\eta'\,dx-=\int_0^1 u_2''\eta dx+u_2'(1)\eta(1)-u_2'(0)\eta(0),$$

or, if we use the conditions (9) which the function u_2 satisfies

$$\int_0^1 u_2'\eta'\,dx = -\int_0^1 u_2''\eta\,dx - u_2(1)\eta(1).$$

Substituting this in the second bracket in (11), we find that this bracket equals

$$-2\int_0^1 (u_2''+2)\eta\,dx = 0,$$

because $u_2''=-2$. Now

$$F(u)=F(u_2)+\left\{\int_0^1 \eta'^2\,dx+\eta^2(1)\right\},$$

and clearly, $F(u)>F(u_0)$ if only $\eta\not\equiv 0$.

The extremely interesting fact disclosed by this example is connected with the fact that conditions (9) arise *naturally* for the functional (8), after its transformation to the form (10). The question of natural boundary conditions will be considered in more detail in § 17.

Theorem 2 proved in this paragraph has a conditional character as it assumes the existence of a solution of equation (1) or (in the second part of theorem 2) the existence of a function imparting a minimum value to the functional $F(u)$. *In the following, in speaking of the equation $Au=f$ where A is a positive definite operator, we shall assume, if nothing is stated to the contrary, that this equation has a solution with a finite energy, i.e. a solution for which the norm of the energy $\|u\|=\sqrt{(Au, u)}$ is finite.*

In § 19 it will be proved that such a solution exists if the operator A is positive-bounded-below and the function f has a finite norm.

§ 12. Representation of a solution in the form of an orthogonal series

Suppose that the equation

$$Au=f(P), \tag{1}$$

where A is a positive definite operator, and the function $f(P)$ possesses a finite norm, has a solution $u_0(P)$ with finite

energy. Then it is possible to construct this solution in the following way:

We select a sequence *which is complete in energy* of functions $\varphi_n(P)$. We shall also assume that this sequence is *energy orthonormal* – this can always be secured by the use, when necessary, of the process of orthogonalization. The function $u_0(P)$ may be expanded into the orthogonal series

$$u_0(P)=\sum_{n=1}^{\infty}a_n\varphi_n(P); \quad a_n=[u_0, \varphi_n].$$

By formula (4) § 8

$$[u_0, \varphi_n]=(Au_0, \varphi_n),$$

u_0 however, satisfies equation (1), and so

$$Au_0=f, \qquad a_n=(f, \varphi_n),$$

and finally

$$u_0(P)=\sum_{n=1}^{\infty}(f, \varphi_n)\varphi_n(P). \tag{2}$$

Series (2) allows us to construct the solution of equation (1), since its coefficients are expressed simply by the known functions $f(P)$ and $\varphi_n(P)$. The series mentioned converges in energy; it also converges in the mean if the operator A is positive-bounded-below.

Example. We consider the problem of the torsion of a rod, the base of which is the rectangle $0\leq x\leq a$, $0\leq y\leq b$. The torsion function $\psi(x, y)$ satisfies the Poisson equation.

$$-\nabla^2\psi=2G\theta, \tag{3}$$

where G is the shear modulus, and θ the angular twist per unit length of rod. Further this function vanishes at the edges of the rectangle $x=0$, $x=a$, $y=0$, $y=b$. As we have seen (§ 6) the operator $-\nabla^2$ is positive definite if its field of definition consists of functions vanishing at the boundary of the region. In our case the energy product of the functions $u(x, y)$ and $v(x, y)$ is expressed by the formula

$$[u, v]=-\int_0^a\int_0^b v(x, y)\nabla^2u(x, y)\mathrm{d}x\mathrm{d}y.$$

We have correspondingly, for the energy norm

$$\|u\|^2 = -\int_0^a \int_0^b u(x, y)\, \nabla^2 u(x, y)\, dx dy.$$

The functions

$$\varphi_{mn}(x, y) = \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b}; \qquad m, n = 1, 2, 3, \ldots \tag{4}$$

are continuously differentiable as many times as required and vanish at the extremities of the rectangle and thus enter into the field of definition of the operator in our problem. They are orthogonal in energy. For a simple proof of this we note firstly that

$$\nabla^2 \varphi_{mn}(x, y) = -\pi^2 \left(\frac{m^2}{a^2} + \frac{n^2}{b^2} \right) \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b}$$

$$= -\pi^2 \left(\frac{m^2}{a^2} + \frac{n^2}{b^2} \right) \varphi_{mn}(x, y).$$

Now

$$[\varphi_{mn}, \varphi_{rs}] = -\int_0^a \int_0^b \varphi_{mn} \nabla^2 \varphi_{rs}\, dx dy$$

$$= \pi^2 \left(\frac{r^2}{a} + \frac{s^2}{b^2} \right) \int_0^a \sin \frac{m\pi x}{a} \sin \frac{r\pi x}{a}\, dx \int_0^b \sin \frac{n\pi y}{b} \sin \frac{s\pi y}{b}\, dy. \tag{5}$$

If at least one of the inequalities $m \neq r$ or $n \neq s$ is true, then $[\varphi_{mn}, \varphi_{rs}] = 0$ which was to be proved. Putting $r = m$ and $s = n$ we find

$$\|\varphi_{mn}\|^2 = \frac{\pi^2 (b^2 m^2 + a^2 n^2)}{4ab},$$

so that the system (5) is not normalized. Dividing $\varphi_{mn}(x, y)$ by $\|\varphi_{mn}\|$ we obtain a system

$$\psi_{mn}(x, y) = \frac{2}{\pi} \sqrt{\left(\frac{ab}{b^2 m^2 + a^2 n^2} \right)} \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b}, \tag{6}$$

which is energy orthonormal.

By formula (2) the torsion function is represented by the series

$$\psi(x, y) = \sum_{m, n=1}^{\infty} (2G\theta, \psi_{mn}) \psi_{mn}(x, y). \tag{7}$$

We now evaluate the coefficients of this series:

$$(2G\theta, \psi_{mn}) = \int_0^a \int_0^b 2G\theta \psi_{mn}(x, y)\, dx dy$$

$$= \frac{4G\theta}{\pi} \left(\frac{ab}{b^2m^2+a^2n^2} \right) \int_0^a \sin \frac{m\pi x}{a} \, dx \int_0^b \sin \frac{n\pi y}{b} \, dy$$

$$= \frac{4abG\theta}{\pi^3 mn} \left(\frac{ab}{b^2m^2+a^2n^2} \right) [1-(-1)^m][1-(-1)^n].$$

It is thus clear that the coefficients of the series (7) equal zero if at least one of the numbers m or n is even; if these numbers are both odd, then

$$(2G\theta, \psi_{mn}) = \frac{16abG\theta}{\pi^3 mn} \left(\frac{ab}{b^2m^2+a^2n^2} \right).$$

Substituting this in (7) we obtain

$$\psi(x, y) = \frac{32G\theta a^2b^2}{\pi^4} \sum_{m, n=1, 3, 5, \ldots} \frac{\sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b}}{mn(b^2m^2+a^2n^2)}. \tag{8}$$

From the general theory it follows that the series (8) converges in energy. As formula (5) § 6 shows, this means that the corresponding series of first derivatives converge in the mean. It can be proved (see §§ 22 and 23) that the operator $-\nabla^2$ in our problem is positive-bounded-below, and so the series (8) converges in the mean. Indeed the series mentioned also converges uniformly.

However, simple, albeit laborious, calculations show that the convergence in the mean of the series (8) and of the series for its derivatives is considerably more rapid than their uniform convergence (see the end of § 7).

§ 13. The minimizing sequence and its convergence

Let $\Phi(u)$ be some functional the values of which are bounded below. Then (cf. 1, 42) there exists an exact lower bound d for the functional $\Phi(u)$:

$$d = \inf \Phi(u).$$

We now introduce a definition: the sequence u_n, $n = 1, 2, \ldots$ of functions belonging to the field of definition of the functional $\Phi(u)$ is known as the *minimizing sequence* if

$$\lim_{n \to \infty} \Phi(u_n) = d.$$

Now let A be a positive definite operator. If the equation

$$Au=f \tag{1}$$

has a solution u_0, then, as we have seen, the functional

$$F(u)=(Au,\ u)-2(u,\ f)\doteq\int_{\Omega}(uAu-2uf)\mathrm{d}\Omega$$

reduces to the form

$$F(u)=\|u-u_0\|^2-\|u_0\|^2. \tag{2}$$

It is therefore clear that $\inf F(u) = \min F(u)=-\|u_0\|^2$. The minimizing sequence for the functional $F(u)$ is characterized by the equation

$$\lim F(u_n)=-\|u_0\|^2. \tag{3}$$

Of great importance for the energy method, is the following statement.

THEOREM 1. *If equation* (1) *has a solution then any sequence which is minimizing for the functional* (2) *converges in energy to this solution.*

The proof is very simple: if the sequence u_n, $n=1, 2, \ldots$ is minimizing for the functional (2), then from (2) and (3) it follows that

$$F(u_n)=\|u_n-u_0\|^2-\|u_0\|^2\underset{n\to\infty}{\longrightarrow}-\|u_0\|^2.$$

Hence $\|u_n-u_0\|\overset{\mathrm{E}}{\to}0$ i. e. $u_n(P)\to u(P)$. Hence, if A is a positive-bounded-below operator, simultaneously $u_n\xrightarrow{\text{mean}}u$.

Theorem 1 underlies many direct methods. Thus, to solve equation (1) (assuming a solution exists) it suffices to construct for the integral (2) a minimizing sequence; any of its terms is an approximate solution of the given equation. At various times many methods of constructing a minimizing sequence have been suggested. One such method was contained essentially in the previous section. If by $u_n(P)$ we denote the partial sums of series (2) § 12:

$$u_n(P)=\sum_{k=1}^{n}(f,\ \varphi_k)\varphi_k(P),$$

then $u_n(P) \xrightarrow{E} u(P)$. The well-known shortcoming of this method is the labour involved in the orthogonalization process. The most important method of constructing a minimizing sequence is the Ritz method; we shall consider also the method of Courant [1], the method due to L. V. Kantorovich, of reduction to a system of ordinary differential equations, (see L. V. Kantorovich and V. I. Krylov [1]), and the method of steepest descent (see L. V. Kantorovich [1] Ch. III).

The details of the methods cited will be expounded to a greater or lesser degree in the next section of the present chapter. We shall mention also the method of least squares to which Chapter IX of this book is devoted. We shall also apply this method to equations of the form (1), where the operator A is non-positive definite, and even asymmetric, but if A is positive definite the method of least squares reduces to a minimizing sequence as will be proved in due course.

§ 14. The Ritz method

The Ritz method is one of the methods of constructing a minimizing sequence. It was mentioned in the introduction in its application to the concrete problems, which were considered by Ritz himself. Here we discuss it for the more general case.

The solution of the equation

$$Au = f(P), \tag{1}$$

where A is a positive definite operator, reduces, as we already know, to the finding of the minimum of the functional:

$$F(u) = (Au, u) - 2(u, f). \tag{2}$$

We shall find an approximate solution to this latter problem in the following manner: We select a sequence

$$\varphi_1(P), \varphi_2(P), \ldots, \varphi_n(P), \ldots \tag{3}$$

of functions belonging to the field of definition D_A of the operator. This sequence is subject to the two conditions:
1) *the sequence* (3) *is complete in energy;*
2) *for any n the functions* $\varphi_1(P)$, $\varphi_2(P)$, ..., $\varphi_n(P)$ *are linearly independent.*

We shall, following Ritz, call functions (3) *coordinate functions*. We construct a linear combination of the first n coordinate functions

$$u_n(P)=\sum_{j=1}^{n} a_j\varphi_j(P) \tag{4}$$

with arbitrary numerical coefficients a_j and substitute $u_n(P)$ for $u(P)$ in functional (2). This makes $F(u)$ into a function of the n independent variables $a_1, a_2, \ldots, a_n$:

$$\begin{aligned} F(u_n)&=\left(\sum_{j=1}^{n} a_jA\varphi_j,\ \sum_{k=1}^{n} a_k\varphi_k\right)-2\left(\sum_{j=1}^{n} a_j\varphi_j,\ f\right)\\ &=\sum_{j,k=1}^{n}(A\varphi_j,\ \varphi_k)a_ja_k-2\sum_{j=1}^{n}(\varphi_j,\ f)a_j. \end{aligned} \tag{5}$$

We select the coefficients a_j so that the function (5) is a minimum. As we shall now see, this condition reduces to a system of linear algebraic equations in the unknowns $a_1, a_2, \ldots, a_n$. Function (5) becomes a minimum for the values of the independent variables which make its first derivatives zero:

$$\frac{\partial F(u_n)}{\partial a_i}=0;\quad i=1,\ 2,\ \ldots,\ n. \tag{6}$$

Equation (6) gives, as is known, the necessary conditions for the minimum of $F(u_n)$. However, by using the fact that the operator A is positive definite it can be shown that the coefficients $a_1, a_2, \ldots, a_n$ which satisfy system (6) realize the minimum magnitude of $F(u_n)$. This remark also holds true for eigenvalue problems (§ 32).

In order to evaluate the derivatives entering into the left-hand side of equation (6), we note that the function $F(u_n)$ is a three-termed quadratic expression in a_i and that by sepa-

rating the terms which contain a_i to the first or second power, we can reduce $F(u_n)$ to the form

$$F(u_n)=(A\varphi_i,\ \varphi_i)a_i^2+\sum_{k\neq i}(A\varphi_i,\ \varphi_k)a_ia_k$$
$$+\sum_{j\neq i}(A\varphi_j,\ \varphi_i)a_ja_i-2(f,\ \varphi_i)a_i+\ \ldots\ ; \tag{7}$$

the dots denoting the group of terms not containing a_i. Expression (7) can be simplified somewhat. First of all it does not matter by which letter the index of summation is denoted, and therefore in the second summation we may write k instead of j. Further since the positive definite operator A is symmetric $(A\varphi_k,\ \varphi_i)=(\varphi_k,\ A\varphi_i)=(A\varphi_i,\ \varphi_k)$, and the second summation in (7) becomes identical with the first. Thus

$$F(u_n)=(A\varphi_i,\ \varphi_i)a_i^2+2\sum_{k\neq i}(A\varphi_i,\ \varphi_k)a_ia_k-2(f,\ \varphi_i)a_i+\ \ldots\ .$$

Now the differentiation is easily performed

$$\frac{\partial F(u_n)}{\partial a_i}=2(A\varphi_i,\ \varphi_i)a_i+2\sum_{k\neq i}(A\varphi_i,\ \varphi_k)a_k-2(f,\ \varphi_i).$$

The first term can be combined with the summation, thus simplifying the equation, and giving

$$\frac{\partial F(u_n)}{\partial a_i}=2\sum_{k=1}^{n}(A\varphi_i,\ \varphi_k)a_k-2(f,\ \varphi_i).$$

Equating the derivatives obtained to zero, we easily arrive at the system of Ritz equations

$$\sum_{k=1}^{n}(A\varphi_i,\ \varphi_k)a_k=(f,\ \varphi_i),\quad i=1,\ 2,\ \ldots,\ n. \tag{8}$$

Writing this system out at greater length, we obtain

$$\begin{aligned}
&(A\varphi_1,\ \varphi_1)a_1+(A\varphi_1,\ \varphi_2)a_2+\ \ldots\ +(A\varphi_1,\ \varphi_n)a_n=(f,\ \varphi_1),\\
&(A\varphi_2,\ \varphi_1)a_1+(A\varphi_2,\ \varphi_2)a_2+\ \ldots\ +(A\varphi_2,\ \varphi_n)a_n=(f,\ \varphi_2),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&(A\varphi_n,\ \varphi_1)a_1+(A\varphi_n,\ \varphi_2)a_2+\ \ldots\ +(A\varphi_n,\ \varphi_n)a_n=(f,\ \varphi_n).
\end{aligned} \tag{8$_1$}$$

It is possible to give the Ritz system of equations another form, if we use the concept of energy product. In this case $(A\varphi_i, \varphi_k)=[\varphi_i, \varphi_k]$ and system (8_1) assumes the form

$$\begin{aligned}
&[\varphi_1, \varphi_1]a_1+[\varphi_1, \varphi_2]a_2+ \ldots +[\varphi_1, \varphi_n]a_n=(f, \varphi_1),\\
&[\varphi_2, \varphi_1]a_1+[\varphi_2, \varphi_2]a_2+ \ldots +[\varphi_2, \varphi_n]a_n=(f, \varphi_2),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&[\varphi_n, \varphi_1]a_1+[\varphi_n, \varphi_2]a_2+ \ldots +[\varphi_n, \varphi_n]a_n=(f, \varphi_n).
\end{aligned} \tag{8_2}$$

The determinant of system (8)

$$\begin{vmatrix}
[\varphi_1, \varphi_1]; & [\varphi_1, \varphi_2]; & \ldots ; & [\varphi_1, \varphi_n]\\
[\varphi_2, \varphi_1]; & [\varphi_2, \varphi_2]; & \ldots ; & [\varphi_2, \varphi_n]\\
\ldots\ldots & \ldots\ldots & \ldots & \ldots\ldots\\
[\varphi_n, \varphi_1]; & [\varphi_n, \varphi_2]; & \ldots ; & [\varphi_n, \varphi_n]
\end{vmatrix} \tag{9}$$

is the Gram determinant of the linearly independent functions $\varphi_1(P)$, $\varphi_2(P)$, ..., $\varphi_n(P)$ and this is different from zero. Hence it follows that the Ritz system of equations is always soluble if the operator A is positive definite. By finding the coefficients a_1, a_2, ..., a_n and substituting them in (4) we obtain the function $u_n(P)$ which we shall call *Ritz's approximate solution of equation* (1).

THEOREM 1. *Ritz's approximate solution of equation* (1) *constitutes a minimizing sequence for the functional* (2) *provided that equation* (1) *has a solution with finite energy.*

For brevity we shall denote the least value of this functional by d:

$$d=\min F(u)=-\|u_0\|^2,$$

where u_0 is the exact solution of equation (1). We note that $F(u_n)\geq d$, i. e. the Ritz method gives a value of the minimal integral which is in error. This immediately follows from the fact that $u_n(P)\in D_A$, $u_n(P)$ being a linear combination of the functions $\varphi_1(P)$, $\varphi_2(P)$, ..., $\varphi_n(P)$ which enter into D_A. The quantity d is the exact lower bound of the functional $F(u)$. If ε is an arbitrary positive number, then, by the very definition of the exact lower bound, there exists a function $v(P)\in D_A$ such that $d\leq F(v)<d+\frac{\varepsilon}{2}$. Further the sequence

φ_n, $n=1, 2, \ldots$ is complete in energy. Hence there follows, as we shall now prove, the possibility of so choosing a natural number n and constants $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that they satisfy the inequality

$$F(v_n)-F(v)<\frac{\varepsilon}{2},$$

where $v_n=\alpha_1\varphi_1+\alpha_2\varphi_2+\ldots+\alpha_n\varphi_n$. In fact, by formula (8) § 11

$$F(v_n)-F(v)=|v_n-u_0|^2-|v-u_0|^2$$
$$=[|v_n-u_0|+|v-u_0|]\,[|v_n-u_0|-|v-u_0|].$$

By the triangle inequality [formula (6) § 8]

$$|v_n-u_0|-|v-u_0|\leq|v_n-v|$$

and

$$F(v_n)-F(v)\leq[|v_n-u_0|+|v-u_0|]\,|v_n-v|.$$

We select n and the coefficients $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that

$$|v_n-v|<\frac{\varepsilon}{k}.$$

The value of k will be chosen later. Now

$$|v_n|<|v|+\frac{\varepsilon}{k}$$

and

$$F(v_n)-F(v)<\left[2|v|+2|u_0|+\frac{\varepsilon}{k}\right]\frac{\varepsilon}{k}.$$

By taking k such that

$$\frac{1}{k}\left[2|v|+2|u_0|+\frac{\varepsilon}{k}\right]<\frac{1}{2},$$

we obtain also the inequality $F(v_n)-F(v)<\frac{\varepsilon}{2}$. Hence it follows that

$$d\leq F(v_n)\leq F(v)+\frac{\varepsilon}{2}<d+\varepsilon.$$

Let u_n be a function constructed by the Ritz method. Then

$$d \leq F(u_n) \leq F(v_n), \quad \text{or} \quad d \leq F(u_n) \leq d + \varepsilon.$$

Letting $\varepsilon \to 0$, we find that $F(u_n) \to d$ i.e., that the sequence u_n, $n = 1, 2, \ldots$ is minimizing.

From the theorem just proved and from the theorem of § 13 about the convergence of a minimizing sequence it follows that Ritz's approximate solutions converge in energy to the exact solution of equation (1). They also converge to the exact solution in the mean if A is a positive-bounded-below operator.

We now explain the connexion between the exact solution of the minimal problem defined by formula (2) § 12 and the approximate solution u_n which the Ritz method gives. Let the functions φ_k be replaced by the functions ψ_k according to the following formulae:

$$\begin{aligned}
&\psi_1 = a_{11}\varphi_1, \\
&\psi_2 = a_{21}\varphi_1 + a_{22}\varphi_2, \\
&\ldots\ldots\ldots\ldots \\
&\psi_n = a_{n1}\varphi_1 + a_{n2}\varphi_2 + \ldots + a_{nn}\varphi_n \\
&\ldots\ldots\ldots\ldots\ldots
\end{aligned}$$

all the constants a_{kk} being different from zero. Expressing u_n in the form $u_n = \sum_{k=1}^{n} b_k \psi_k$ or in the form $u_n = \sum_{k=1}^{n} a_k \varphi_k$, we arrive at one and the same result, since the expression

$$\sum_{k=1}^{n} b_k \psi_k = \sum_{k=1}^{n} b_k \sum_{m=1}^{k} a_{km} \varphi_m = \sum_{m=1}^{n} \varphi_m \sum_{k=m}^{n} a_{km} b_k$$

is also a linear combination of the functions φ_k. Bearing this in mind, we subject the sequence $\{\varphi_n\}$ to energy orthogonalization. To avoid the use of extra symbols we shall denote the new functions which we obtained by the orthogonalization process by φ_k as before. Now

$$(A\varphi_k, \varphi_j) = [\varphi_k, \varphi_j] = \begin{cases} 0, & j \neq k \\ 1, & j = k \end{cases} \tag{10}$$

and system (8) assumes the particularly simple form

$$a_j=(f,\ \varphi_j),\quad j=1,\ 2,\ \ldots,\ n.$$

Now

$$u_n=\sum_{k=1}^{n}(f,\ \varphi_k)\varphi_k, \tag{11}$$

i.e. the approximate solution of the minimal problem obtained by the Ritz method coincides with cutting short the series (2) § 12 which represents the exact solution.

Ritz's approximate solution u_n is nearest (in the energy sense) to the exact one the larger the value of n. To secure high accuracy it is necessary to take n, i.e. a large number of coordinate functions. This leads to the necessity of solving system (8) with a large number of equations and unknowns. In order to become acquainted with the practical methods of oslving such systems we refer the reader to the book by V. N. Faddeyeva [4]. We merely note that the solution of system (8) si facilitated by its symmetry. This means that the coefficients $[\varphi_i,\ \varphi_k]$ and $[\varphi_k,\ \varphi_i]$ which are symmetric with respect to the principal diagonal are equal.

N O T E: If system (3) has already been constructed and it is desirable for some reason or other to construct a less accurate approximation which does not contain several of the coordinate functions, then the coefficients a_k corresponding to this less accurate approximation are found from the system which is obtained by truncating system (8), i. e. by striking out the rows and columns corresponding to the coordinate functions that have been rejected. A similar remark is also valid for the problem of eigenvalues (see § 32).

Let, as before $\varphi_n(P)$, $n=1,\ 2,\ \ldots$ denote coordinate functions, energy orthonormalized, such that relationship (10) holds.

Putting $c_i=(f,\ \varphi_i)$ for brevity, we have

$$u_n=\sum_{i=1}^{n}c_i\varphi_i,\qquad u_0=\sum_{i=1}^{\infty}c_i\varphi_i.$$

We have already seen (formula (3) § 11) that $F(u_0) = -\|u_0\|^2$ and we shall prove that if $u_n(P)$ is an approximate solution of equation (1) constructed by the Ritz method, then

$$F(u_n) = -\|u_n\|^2. \tag{12}$$

In fact $F(u_n) = \|u_n\|^2 - 2(u_n, f)$, but

$$(u_n, f) = \left(\sum_{i=1}^{n} c_i \varphi_i, f\right) = \sum_{i=1}^{n} c_i (\varphi_i, f) = \sum_{i=1}^{n} c_i^2.$$

Further, the functions $\varphi_i(P)$ are energy orthonormal. From what was proved in § 10 [formulae (20) and (21)].

$$\|u_n\|^2 = \sum_{i=1}^{n} c_i^2.$$

Hence it follows that

$$(u_n, f) = \|u_n\|^2 \tag{13}$$

and $F(u_n) = -\|u_n\|^2$. Passing to the limit as $n \to \infty$ in (13) we also find that

$$(u_0, f) = \|u_0\|^2. \tag{14}$$

This, however, is also immediately obvious, since

$$(u_0, f) = (u_0, Au_0) = [u_0, u_0] = \|u_0\|^2.$$

From formulae (20) and (21) § 10 it follows that

$$\|u_0\|^2 = \sum_{i=1}^{\infty} c_i^2$$

whence it is clear that as n increases, the quantity $\|u_n\|$ increases, and

$$\|u_n\| \leq \|u_0\|; \quad \|u_n\| \underset{n \to \infty}{\to} \|u_0\|. \tag{15}$$

Let $n > k$. We evaluate the quantity $\|u_n - u_k\|^2$. We have

$$u_n - u_k = \sum_{i=1}^{n} c_i \varphi_i - \sum_{i=1}^{k} c_i \varphi_i = \sum_{i=k+1}^{n} c_i \varphi_i;$$

by formulae (20) and (21) § 10

$$\|u_n - u_k\|^2 = \sum_{i=k+1}^{n} c_i^2$$

or

$$\|u_n - u_k\|^2 = \|u_n\|^2 - \|u_k\|^2. \tag{16}$$

Similarly

$$\|u_0 - u_k\|^2 = \|u_0\|^2 - \|u_k\|^2. \tag{17}$$

Let $u(P)$ be an arbitrary function, not identically zero, from the field of definition of the operator A. The function $\varphi_1(P) = \dfrac{u(P)}{\|u\|}$ is energy normalized. We shall consider $\varphi_1(P)$ as the first term of some energy orthonormal system of functions $\varphi_n(P)$ and using the Ritz method we shall seek an approximate solution $u_1(P)$ of equation (1) by putting $n=1$ in formula (4). Then, as we know, $u_1 = c_1\varphi_1$ where $c_1 = (f, \varphi_1)$ and

$$\|u_0\|^2 \geq \|u_1\|^2 = c_1^2.$$

But

$$c_1 = (f, \varphi_1) = \frac{(f, u)}{\|u\|},$$

and so we arrive at the useful inequality

$$\|u_0\|^2 \geq \frac{(f, u)^2}{\|u\|^2}. \tag{18}$$

In inequality (18) $u_0(P)$ means the exact solution of equation (1) and $u(P)$ – any function not identically zero which belongs to the field of definition of the operator A. This means, as we remember, that this function has the necessary number of continuous derivatives and satisfies the boundary conditions of the problem.

We now consider the question of to what extent the stipulations of completeness and linear independence that we have imposed upon the coordinate functions are necessary. The first stipulation is not entirely necessary. In fact there is no necessity that every function from D_A can be approximated in the energy sense by a linear combination of coordinate functions, it being sufficient that the solution sought permits such an approximation. In the proof of theorem 1 it is then

possible to take for v an approximate linear combination of coordinate functions and the whole proof holds true. Therefore if it is known beforehand that the solution sought belongs to a certain class which is more restricted than D_A then it suffices that the system of coordinate functions be complete in this class. For example, if it is known that the function sought is even in a certain independent variable then we may take for the coordinate functions only even functions in that variable, and it suffices that the selected system of coordinate functions be complete with respect to even functions from D_A.

We now explain what would happen if the completeness condition were satisfied, but the coordinate functions were linearly dependent. We shall suppose, for simplicity, that there is only one function in system (3) which is linearly expressed in terms of a finite number of other functions of this system, and that this function is identically different from zero. In this case it is always possible, by substituting for certain coordinate functions their linear combinations to satisfy ourselves that $\varphi_1(P) \equiv 0$ and that the functions $\varphi_2(P)$, $\varphi_3(P)$, ... taken any finite number at a time, would be linearly independent. Expression (4) can be written in the form

$$u_n(P) = \sum_{i=2}^{n} a_j \varphi_j(P),$$

and all the remainder of the analysis follows without alteration. In particular, the coefficients $a_2, a_3, \ldots, a_n$ are uniquely defined. Hence it follows that the coefficients $a_1, a_2, \ldots, a_n$ are also defined but no longer uniquely: the coefficient a_1 can be specified arbitrarily. This means that system (8) is soluble, although not in unique fashion, and any of its solutions leads to the same function $u_n(P)$. Thus, the presence in a system of coordinate functions of one or several functions which are linearly dependent upon the others does not alter the rule for constructing Ritz's approximate solution, although in this case the solution of the Ritz system of equations is more complex. In practical calculations cases are encountered when

the coordinate functions are "quasi-linearly dependent", i. e., when determinant (9) is nearly zero while the coordinate functions, strictly speaking, are linearly independent. In such a case the relative error in the evaluation of the determinant and, with it also, the numerical values of the coefficients $a_1, a_2, \ldots, a_n$ fluctuate markedly with the degree of accuracy of the coefficients and of the independent terms of equation (8), but these fluctuations have little effect upon the function $u_n(P)$.

Thus the presence of linearly dependent or of quasi-linearly dependent functions is no obstacle to the use of the Ritz method. It should, however, be borne in mind that this unnecessarily complicates the nature and solution of the Ritz system of equations, and therefore such quasi or simply linearly dependent functions should be excluded from the system of coordinate functions.

§ 15. Other methods of constructing the minimizing sequence

The Ritz method is the most important but not the only method of constructing the minimizing sequence. In the present paragraph we shall briefly consider three more such methods; one of them given by R. Courant and the other two by L. V. Kantorovich (see L. V. Kantorovich and V. I. Krylov [1] and L.V. Kantorovich [1]). See also the work of M. Sh. Birman [1, 2].

Courant's method. R. Courant [1] has presented a method of constructing the minimizing sequence which together with the sequences associated with the derivatives up to a certain order converges under known conditions, not only in the mean but also uniformly.

Consider the differential equation

$$Au = f. \tag{1}$$

It is required to find its solution defined in a certain finite domain Ω of m-dimensional space $(x_1, x_2, \ldots, x_m)$ and satisfy-

ing on the boundary S of this domain certain homogeneous boundary conditions. We suppose that the operator A is positive definite for the set of functions which satisfy the boundary conditions of the problem, and that the problem has a solution possessing a finite energy. This solution gives the minimum value of the functional

$$F(u)=(Au,\ u)-2(u,\ f).$$

We now assume that the given function f has continuous derivatives in $x_1, x_2, \ldots, x_m$ up to a certain order k inclusive. We construct the new functional

$$\Phi(u)=F(u)+\sum_{j=0}^{k}\sum_{\alpha_1+\alpha_2+\ldots+\alpha_m=j}\left\|\frac{\partial^j(Au-f)}{\partial x_1^{\alpha_1}\partial x_2^{\alpha_2}\ldots\partial x_m^{\alpha_m}}\right\|^2. \quad (2)$$

Clearly $\Phi(u)\geq F(u)$. Furthermore, the function n which gives the minimum of $F(u)$ also gives the minimum of $\Phi(u)$ since this function minimizes both the first term in (2) and the second (which it causes to vanish). Hence it follows that the solution of our boundary problem can be found by finding the solution of the problem of the minimum of $\Phi(u)$. We construct, for example by the Ritz method, a minimizing sequence $\{u_n\}$ for this functional.
Then obviously

$$\left\|\frac{\partial^j(Au_n-f)}{\partial x_1^{\alpha_1}\partial x_2^{\alpha_2}\ldots\partial x_m^{\alpha_m}}\right\|\underset{n\to\infty}{\longrightarrow}0, \qquad j=1,\ 2,\ \ldots,k. \quad (3)$$

Relationship (3) allows us to draw the supplementary conclusions about the convergence of the minimal sequence.

We consider as an example the Poisson equation

$$-\nabla^2 u=f$$

in a plane under the boundary condition $u|_s=0$. In (2) we put $k=0$.
Then

$$\Phi(u)=(-\nabla^2 u,\ u)-2(u,\ f)+\|\nabla^2 u+f\|^2$$

$$= \int_{\Omega} \left\{ \left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 - 2uf + (\nabla^2 u + f)^2 \right\} \mathrm{d}\Omega.$$

We construct the minimizing sequence u_n. Then clearly

$$\int_{\Omega} (\nabla^2 u_n + f)^2 \mathrm{d}\Omega \to 0. \tag{4}$$

By Green's formula,

$$u_n(x, y) - u(x, y) = \int_{\Omega} G(x, y; \xi, \eta)(\nabla^2 u_n + f)\mathrm{d}\Omega,$$

where $G(x, y; \xi, \eta)$ is Green's function for the Laplacian operator.

Hence, by Cauchy's inequality,

$$|u_n(x, y) - u(x, y)| \leq \sqrt{\left(\int_{\Omega} G^2(x, y; \xi, \eta)\mathrm{d}\xi\mathrm{d}\eta \right)} \times \sqrt{\left(\int_{\Omega} (\nabla^2 u_n + f)^2 \mathrm{d}\xi\mathrm{d}\eta \right)}. \tag{5}$$

The Green's function satisfies the inequality

$$0 \leq G(x, y; \xi, \eta) \leq a|\ln r| + b, \tag{6}$$

where a and b are constants (see e. g. V. I. Smirnov [2] 220–224). It is readily noticed that the first factor on the right in (5) is bounded:

$$\sqrt{\left(\int_{\Omega} G^2(x, y; \xi, \eta)\mathrm{d}\xi\mathrm{d}\eta \right)} < C, \qquad C = \text{const.}$$

Now from (5) it follows that $u_n(x, y) \to u(x, y)$ *uniformly* in $\bar{\Omega} = \Omega + S$.

The introduction of additional terms in $\Phi(u)$ complicates the calculations. This complication, however, can be justified if it is desired to obtain a uniformly converging sequence associated with the problem.

K. N. Shevchenko [1] has applied Courant's method to the three-dimensional static problem of elasticity theory.

Kantorovich's method of reduction to ordinary differential equations

L. V. Kantorovich has evolved a method of solving minimal problems which is essentially different from the Ritz method. This method is discussed in detail in the monograph by L. V. Kantorovich and V. I. Krylov [1]; we shall restrict ourselves here to a few remarks only.

For simplicity we assume that the domain Ω is plane and has the form depicted in Fig. 6, and that it is required to

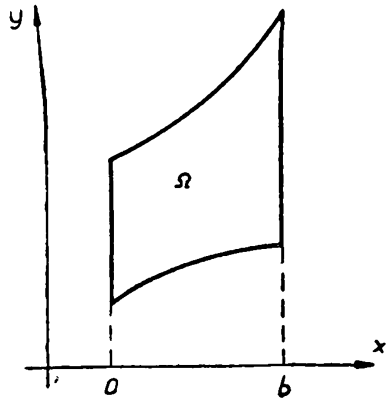

FIG. 6

integrate the differential equation (1) under the condition $u|_s=0$. Let the operator A be positive definite for the set of functions satisfying the given boundary condition. Then our problem reduces to the problem of finding the minimum of the functional $F(u)=(Au, u)-2(u, f)$.

We put

$$u_n(x, y)=\sum_{k=1}^{n}\chi_k(x, y)f_k(x),$$

where $\chi_k(x, y)$ are known functions, equal to zero on S except perhaps on the straight lines $x=a$ and $x=b$. We determine functions $f_k(x)$ of one variable such that the functional $F(u_n)$ has a minimal value. By the usual methods of variational calculus, see e. g. V. I. Smirnov [2], Ch.II, we obtain for $f_k(x)$

a system of differential equations. With these are associated the boundary conditions at $x=a$ and $x=b$ which emerge from the boundary conditions of the problem:

$$f_k(a)=f_k(b)=0, \quad k=1, 2, \ldots, n.$$

We see that the essence of L.V. Kantorovich's method consists of the fact that it approximates the integration of a partial differential equation by the integration of a system of ordinary differential equations.

The method of steepest descent

This method we shall apply directly to those symmetric operators which satisfy the double inequality

$$m\|u\|^2 \leq (Au, u) \leq M\|u\|^2, \tag{7}$$

where m and M are positive constants. Such operators belong to the class of so called *bounded* operators. (For more details about bounded operators see below § 45.) Differential operators do not enter into this class, and so the application of the method of steepest descent to a differential equation requires preliminary transformation of this equation so that the operator on its left-hand side becomes bounded.

Let operator A satisfy inequality (7). From this inequality it follows, *inter alia*, that A is a positive-bounded-below operator. We consider the equation

$$Au=f, \tag{8}$$

where u is the sought and f the given function. By the minimum functional theorem (§ 11) the solution of this equation is equivalent to the solution of the problem of finding a minimum to the functional:

$$F(u)=(Au, u)-2(u, f).$$

We take an arbitrary function u_1. It may happen that $Au_1=f$ then the problem is solved. If $Au_1 \neq f$ then we take u_1 as a first approximation to the solution sought. To construct the second approximation we put $Au_1-f=v_1$ and select a number

a such that the expression $F(u_1 - av_1)$ assumes the least value. Expanding the brackets we easily find

$$F(u_1 - av_1) = F(u_1) - 2a[(Au_1, v_1) - (f, v_1)] + a^2(Av_1, v_1).$$

Equating to zero the first derivative of this expression with respect to a we obtain

$$(Av_1, v_1)a - [(Au_1, v_1) - (f, v_1)] = 0.$$

This equation can be simplified somewhat since

$$(Au_1, v_1) - (f, v_1) = (Au_1 - f, v_1) = (v_1, v_1) = \|v_1\|^2.$$

Now

$$(Av_1, v_1)a - \|v_1\|^2 = 0 \text{ and } a = \frac{\|v_1\|^2}{(Av_1, v_1)} = a_1.$$

As the second approximation we take the function $u_2 = u_1 - a_1 v_1$. By repeating the process we obtain the third approximation u_3 and so on. The rapidity of convergence of the method of steepest descent is estimated by the inequality

$$\|u_{n+1} - u_0\| \leq \|u_1 - u_0\| \left(\frac{M-m}{M+m}\right)^n,$$

where u_0 is the exact solution of equation (8).

§ 16. Functions with finite energy

Let A be a positive operator and $u(P)$ be a function from its field of definition D_A. The quantity $\|u\|^2 = (Au, u)$ is obviously finite and it can be said that any function from the field D_A has a finite energy. In the present paragraph we shall show that the concept of energy can be extended for a wider class of functions than D_A which we shall call *the class of functions with finite energy*. Bearing in mind the objective of this book we shall restrict ourselves to cases where the operator A is positive-bounded-below.

The construction of the class of functions with finite energy is founded upon the well-known *Riesz–Fischer theorem*. (*For*

the proof of this theorem see, e.g. I. P. Natanson [2].) *Let the functions* $\varphi_n(P)$, $n=1, 2, \ldots$ *be defined in some domain and let their norms in this domain be finite. If these functions satisfy the condition*

$$\lim_{n,\, m\to\infty} \|\varphi_n-\varphi_m\| = \lim_{n,\, m\to\infty} \left\{\int_\Omega [\varphi_n(P)-\varphi_m(P)]^2 \mathrm{d}\Omega\right\}^{\frac{1}{2}} = 0, \qquad (1)$$

then there exists a function $\varphi(P)$ *with a finite norm which is the limit, in the sense of convergence in the mean, of the sequence of functions* $\varphi_n(P)$. (*We note that condition* (1) *is similar to that used in the Cauchy test for the existence of a limit* (*cf* **1**,**31**.))

The Riesz–Fischer theorem asserts consequently that the integral

$$\int_\Omega \varphi^2(P)\mathrm{d}\Omega$$

exists and that $\varphi_n(P) \xrightarrow{\text{mean}} \varphi(P)$ or what is the same thing,

$$\lim_{n\to\infty} \int_\Omega [\varphi(P)-\varphi_n(P)]^2 \mathrm{d}\Omega = 0.$$

It should be noted that the function $\varphi(P)$ may not be integrable in the Riemann sense (cf. **1**, **116**) and then the latter two integrals should be understood in the Lebesque sense (see § 41).

Let A be a positive-bounded-below operator so that for any function the inequality

$$\|u\|^2 = (Au,\, u) \geq \gamma^2 \|u\|^2 \qquad (2)$$

is satisfied.

We consider some sequence of functions $u_n(P)$ belonging to the field of definition D_A of the operator A and satisfying the condition

$$\lim_{n,\, m\to\infty} \|u_n-u_m\| = 0. \qquad (3)$$

If the functions $u_n(P)$ and $u_m(P)$ belong to D_A then their difference $u_n(P)-u_m(P)$ also belongs to D_A. From (2) and (3) it follows that

$$\lim_{n,\, m\to\infty} \|u_n-u_m\| = 0,$$

and that, in virtue of the Riesz–Fischer theorem, there exists a function $u(P)$ with a finite norm such that

$$u_n(P) \overset{\text{mean}}{\longrightarrow} u(P). \tag{4}$$

We take functions of the type $u(P)$ in the class of functions with finite energy. It is easily seen that any function belonging to the field of definition D_A of the given operator also belongs to the class of functions with finite energy. In fact if $u \in D_A$ then it suffices to put $u_1 = u_2 = u_3 = \ldots = u$ and then obviously

$$|u_n - u_m| \to 0 \quad \text{and} \quad \|u_n - u\| \to 0.$$

We now show how it is possible to define the energy product and energy norm for functions with finite energy. If $u(P)$ and $v(P)$ are functions with finite energy then by the definition there exist two sequences $u_n(P) \in D_A$ and $v_n(P) \in D_A$ such that

$$\lim_{n,\, m \to \infty} |u_n - u_m| = 0, \qquad \lim_{n,\, m \to \infty} |v_n - v_m| = 0,$$

$$\lim_{n \to \infty} \|u_n - u\| = 0, \qquad \lim_{n \to \infty} \|v_n - v\| = 0.$$

By the triangle inequality for the energy norm (§ 8),

$$\big||u_n| - |u_m|\big| \leq |u_n - u_m| \xrightarrow[n,\, m \to \infty]{} 0.$$

This means that the numerical sequence $|u_n|$, $n = 1, 2, \ldots$ satisfies the Cauchy test for the existence of a limit (cf, **31**), and thus there exists a limit

$$\lim_{n \to \infty} |u_n|.$$

We adopt as a definition

$$|u| = \lim_{n \to \infty} |u_n|, \qquad u_n \in D_A. \tag{5}$$

Similarly

$$|v| = \lim_{n \to \infty} |v_n|, \qquad v_n \in D_A.$$

Substituting $u_n(P)$ in the inequality (2)

$$|u_n|^2 \geq \gamma^2 \|u_n\|^2$$

and let $n \to \infty$. We again arrive at inequality (2) which is thus valid for any function with finite energy.

We construct the energy product of the functions $u_n(P)$ and $v_n(P)$

$$[u_n, v_n] = (Au_n, v_n).$$

We shall prove that as $n \to \infty$ it tends to some limit. It suffices to show that it converges by the Cauchy test mentioned above.

We have:

$$\begin{aligned}[u_n, v_n] - [u_m, v_m] &= [u_n, v_n] - [u_n, v_m] + [u_n, v_m] - [u_m, v_m] \\ &= [u_n, v_n - v_m] + [u_n - u_m, v_m].\end{aligned}$$

By Cauchy's inequality (§ 8)

$$|[u_n, v_n] - [u_m, v_m]| \leq |u_n||v_n - v_m| + |v_m||u_n - u_m|.$$

On the right the first factors are bounded being variables which possess a limit (cf. 1, 27) and the second factors tend to zero as $m \to \infty$ and $n \to \infty$ by the convergence condition. Hence it follows that

$$\lim_{n, m \to \infty} |[u_n, v_n] - [u_m, v_m]| = 0.$$

We apply the Cauchy test to the sequence $[u_n, v_n]$, $n = 1, 2, \ldots$ and thus,

$$\lim_{n \to \infty} [u_n, v_n].$$

exists.

We now define the energy product of functions with finite energy, $u(P)$ and $v(P)$ by putting

$$[u, v] = \lim_{n \to \infty} [u_n, v_n]; \qquad u_n \in D_A, \quad v_n \in D_A. \tag{6}$$

Obviously, all the properties of energy norm and energy product established in § 8 for functions from the field of definition of the operator A hold for functions with finite energy.

Let, as before, $u(P)$ be a function with finite energy and $u_n(P)$, $n = 1, 2, \ldots$ be a sequence of functions from D_A satisfying (3) and (4). We shall prove that

$$u_n(P) \xrightarrow{E} u(P). \tag{7}$$

For the proof we consider the function $P(w)=u(P)-u_k(P)$ where the number k is fixed. It is easily seen that $w(P)$ is a function with finite energy and that this function is connected by relationships of the form (3) and (4) with the sequence of functions $u_n(P)-u_k(P)=\omega(P)$; $n=1, 2, \ldots$. As was stated above

$$\lim_{n\to\infty} |w_n| = |w|,$$

or what is the same thing

$$\lim_{n\to\infty} |u_n-u_k| = |u-u_k|. \tag{8}$$

We now turn to equation (3). Here the passing to the limit can be performed by tending n and m to infinity according to any law. We carry out this transition thus: firstly we arbitrarily fix m and let $n\to\infty$ and then let $m\to\infty$. Then equality (3) assumes the form

$$\lim_{m\to\infty} \{ \lim_{n\to\infty} |u_n-u_m| \} = 0.$$

In virtue of relationship (8) the inner limit equals $|u-u_m|$ and we find

$$\lim_{m\to\infty} |u-u_m| = 0,$$

which is identical with (7).

It is important to note that whether a given function belongs or does not belong to the class of functions with finite energy is determined not only by the given function but also by the given operator A.

We consider as an example the operator

$$Bu = -\frac{d^2u}{dx^2}, \quad 0<x<1. \tag{9}$$

Let its field of definition be characterized by the boundary conditions

$$u(0)=u(1)=0. \tag{10}$$

In § 6 it was established that the operator B is positive-bounded-below and that

$$\|u\|^2=(Bu,\ u)=\int_0^1 u'^2(x)\mathrm{d}x. \tag{11}$$

Let $u_n(x)\in D_B$ and

$$\|u_n-u_m\|^2=\int_0^1[u_n'(x)-u_m'(x)]^2\mathrm{d}x \xrightarrow[n,\,m\to\infty]{} 0. \tag{12}$$

From what has been said above there exists a function $u(x)$ with finite energy such that $u_n(x)\xrightarrow{\text{mean}}u(x)$ and $u_n(x)\xrightarrow{\text{E}}u(x)$. We now attempt to elucidate the properties of the function $u(x)$. As was stated previously (§ 4) we assume that functions from the field of definition D_B of the operator B possess continuous second derivatives. From formula (12) and from the Riesz–Fischer theorem it follows that the first derivatives $u_n'(x)$ converge in the mean to some function which we denote by $v(x)$. If the function $u(x)$ has a continuous derivative then, as is easily proved, $v(x)=u'(x)$. In the general case, corresponding with the definition given by S. L. Sobolev [2], $v(x)$ is called the *generalized derivative* of the function $u(x)$. Thus the function $u(x)$ which possesses finite energy (with respect to the operator defined by (9) and (10)) in the general case has only a generalized first derivative. We shall prove that $u(x)$ is continuous and satisfies conditions (10).

We have

$$u_n(x)-u_m(x)=u_n(0)-u_m(0)+\int_0^x[u_n'(t)-u_m'(t)]\mathrm{d}t.$$

The function $u_n(x)$ satisfies conditions (10) since it belongs to D_B.

Thus $u_n(0)-u_m(0)=0$ and

$$u_n(x)-u_m(x)=\int_0^x[u_n'(t)-u_m'(t)]\mathrm{d}t.$$

By Cauchy's inequality

$$|u_n(x)-u_m(x)| \leq \left\{x\int_0[u_n'(t)-u_m'(t)]^2\mathrm{d}t\right\}^{\frac{1}{2}}$$

$$\leq \left\{\int_0^1[u_n'(t)-u_m'(t)]^2\mathrm{d}t\right\}^{\frac{1}{2}}. \qquad (13)$$

The first part of this inequality does not depend upon x and tends to zero as $n, m\to\infty$. By the Weierstrass test (cf. 1, **147**) the sequence converges uniformly to some continuous function $\bar{u}(x)$.

It is easily seen that $\bar{u}(x)=u(x)$. In fact convergence in the mean follows from uniform convergence (§ 7). $u_n(x)\xrightarrow{\text{mean}}\bar{u}(x)$ but the same sequence cannot converge in the mean to two different limits (theorem 1 § 7) and thus $\bar{u}(x)=u(x)$ and $u_n(x)\to u(x)$ uniformly in the interval $0\leq x\leq 1$. In particular

$$u(0)=\lim_{n\to\infty}u_n(0)=0; \qquad u(1)=\lim_{n\to\infty}u_n(1)=0.$$

Thus, in the example considered functions with finite energy satisfy the same boundary conditions as functions from the field of definition of the given operator; the essential difference lies only in the fact that functions with finite energy need not have continuous derivatives.

We consider another operator

$$B_0u=-\frac{\mathrm{d}^2u}{\mathrm{d}x^2},\ 0<x<1;\ u'(0)=u'(1)=0. \qquad (14)$$

It is easily seen that this operator is not positive definite. In fact

$$(B_0u,\ u)=-\int_0 u\frac{\mathrm{d}^2u}{\mathrm{d}x^2}\mathrm{d}x=-[uu']_0^1+\int_0 u'^2(x)\mathrm{d}x=\int_0^1 u'^2(x)\mathrm{d}x. \qquad (15)$$

The terms integrated out vanish by virtue of the boundary conditions. Formula (15) shows that $(B_0u,\ u)\geq 0$. On the other

hand the function $u(x)\equiv 1$ has continuous derivatives and satisfies boundary conditions (14); at the same time $(B_0 1, 1)= =0$ which contradicts the definition of an operator being positive definite.

In order to make B_0 a positive definite operator, we restrict its field of definition. For this purpose we impose upon the functions of this field, besides boundary conditions (14), the additional condition

$$\int_0^1 u(x)\mathrm{d}x=0. \tag{16}$$

With this additional condition the operator B_0 becomes positive definite and in fact positive-bounded-below. We now prove this. Let x_1 and x_2 be arbitrary points in the interval $0\leq x\leq 1$ and let $u(x)$ be an arbitrary function with a continuous derivative over this interval. We have

$$u(x_1)-u(x_2)=\int_{x_2}^{x_1} u'(t)\mathrm{d}t.$$

Whence, if $x_2\leq x_1$

$$[u(x_1)-u(x_2)]^2\leq\left(\int_{x_2}^{x_1} u'(t)\mathrm{d}t\right)^2$$

$$\leq(x_1-x_2)\int_{x_2}^{x_1} u'^2(t)\mathrm{d}t\leq\int_0^1 u'^2(t)\mathrm{d}t;$$

the same inequality is obtained if $x_2>x_1$. We remove the bracket on the left:

$$u^2(x_1)+u^2(x_2)-2u(x_1)u(x_2)\leq\int_0^1 u'^2(t)\mathrm{d}t.$$

We integrate the inequality obtained with respect to x_1 and x_2 over the range for both from zero to unity. This leads us to the inequality

$$2\int_0^1 u^2(x)\,\mathrm{d}x-2\left(\int_0^1 u(x)\mathrm{d}x\right)^2\leq\int_0^1 u'^2(t)\mathrm{d}t. \tag{17}$$

Now let $u(x) \in D_{B_0}$. In virtue of condition (16) the second integral on the left in (17) vanishes and we obtain

$$\int_0^1 u'^2(x)\mathrm{d}x \geq 2\int_0^1 u^2(x)\mathrm{d}x = 2\|u\|^2,$$

which in virtue of (15) can be put in the form

$$(B_0 u,\ u) \geq 2\|u\|^2.$$

This inequality proves that B_0 is a positive-bounded-below operator. We shall now prove that for the operator B_0 the class of functions with finite energy also contains functions which do not satisfy boundary conditions (14). Let, for example, $u(x)$ be any function which is continuous and continuously differentiable in $0 \leq x \leq 1$ and satisfies equation (6) but does not satisfy condition (14). As is known from the theory of Fourier series (cf. 2, **145**) the functions $\sin k\pi x$ $k=1, 2, \ldots$, constitute a complete orthogonal system in the interval $0 \leq x \leq 1$. The derivative $u'(x)$ can be expanded into the series

$$u'(x) = \sum_{k=1}^{\infty} a_k \sin k\pi x;\quad a_k = 2\int_0^1 u'(x)\sin k\pi x\mathrm{d}x, \tag{18}$$

which in each case converges in the mean (§ 7) in the interval $0 \leq x \leq 1$. We integrate this series term by term in the interval from zero to x, which is admissible by virtue of theorem 2 § 7.

$$u(x) = C - \sum_{k=1}^{\infty} \frac{a_k}{\pi k} \cos k\pi x.$$

Condition (16) gives $C=0$ and finally,

$$u(x) = -\sum_{k=1}^{\infty} \frac{a_k}{\pi k} \cos k\pi x. \tag{19}$$

We now put

$$u_n(x) = -\sum_{k=1}^{n} \frac{a_k}{\pi k} \cos k\pi x.$$

The functions $u_n(x)$ are continuous and continuously differentiable as many times as required. Moreover, they satisfy conditions (14) since

$$u_n'(x) = \sum_{k=1}^{n} a_k \sin k\pi x$$

and thus $u_n'(0) = u_n'(1) = 0$ and $u_n(x) \in D_{B_0}$.

From expansion (18), which in every case converges in the mean, it follows that

$$\|u_n'(x) - u_m'(x)\| = \|u_n - u_m\| \xrightarrow[n, m \to \infty]{} 0.$$

By definition there exists a function $\bar{u}(x)$ with finite energy to which the sequence $u_n(x)$ converges both in energy and in the mean. From expansion (19) it follows that $u_n(x) \xrightarrow{\text{mean}} u(x)$. But then $\bar{u}(x) = u(x)$ and this means that $u(x)$ is a function with finite energy. This function does not, however, satisfy boundary conditions (14).

§ 17. The applications of functions with finite energy. Natural boundary conditions

We consider a positive-bounded-below operator A, the equation $Au = f(P)$ and the corresponding variational problem of finding the minimum of the functional

$$F(u) = (Au, u) - 2(u, f).$$

It is obvious that the functional $F(u)$ is defined only in the case when $u \in D_A$ as only then does the expression (Au, u) have a meaning. However, if we make use of the concepts of energy product and energy norm then we can write $(Au, u) = \|u\|^2 = [u, u]$ and give our functional the new form

$$F(u) = [u, u] - 2(u, f). \tag{1}$$

In such a form the functional $F(u)$ is defined for a wider class of functions: *the right-hand side of equation* (1) *has meaning if* $u(P)$ *is any function with finite energy.*

In § 11 the problem of the minimum of the functional $F(u)$ in the class of functions constituting the field of definition D_A of the operator A was considered. It was explained that the function $u_0(P)$ which satisfies the equation $Au_0=f(P)$ gives the minimum value. We now pose the problem of finding the minimum of the same functional in the wider class of *functions with finite energy*. We shall prove that in this case the minimum of the functional is not altered and that the function $u_0(P)$ only gives the minimum value.

We denote by d the minimum of $F(u)$ in the class D_A and by d_1 the minimum of $F(u)$ in the class of functions of finite energy. Obviously in extending the class of functions the minimum functional cannot increase so $d_1 \leq d$. We suppose that $d_1 < d$.

Then we must find a function $\tilde{u}(P)$ with finite energy such that $F(\tilde{u}) < d$ or, in full

$$[\tilde{u}, \tilde{u}] - 2(\tilde{u}, f) = |\tilde{u}|^2 - 2(\tilde{u}, f) < d.$$

Corresponding with formulae (4) and (7) § 16 there exists a sequence of functions $u_n \in D_A$ such that $|u_n - \tilde{u}| \to 0$, $\|u_n - \tilde{u}\| \to 0$. From the first relationship it follows that $|u_n| \to |\tilde{u}|$ and from the second that $(u_n, f) \to (u, f)$ (see § 3). But then for sufficiently large n the quantities $F(\tilde{u}) = |\tilde{u}|^2 - 2\ (u, f)$ and $F(u_n) = |u_n|^2 - 2\ (u, f)$ differ by as little as we please, and thus if the number n is sufficiently large, then $F(u_n) < d$. This, however, is impossible, since $u_n \in D_A$ and d is the least of the values which the functional $F(u)$ assumes for functions of the class D_A. The contradiction obtained shows that $d_1 = d$.

We suppose that in the class of functions with finite energy the minimum value of the functional $F(u)$ is given not only by the function $u_0(P)$ mentioned above but also by some other function $w_0(P)$. We shall prove that $w_0(P) \equiv u_0(P)$. Formula (5) § 11 gives an identity valid for any function $\eta\ (P)$ of the class D_A: $(Au_0, \eta) = (f, \eta)$ or, if we use the concept of energy product

$$[u_0, \eta] = (f, \eta). \tag{2}$$

Now putting $\eta = u_0$ we obtain, in particular

$$[u_0, u_0] = (f, u_0). \tag{3}$$

We can easily convince ourselves that identity (2) holds true for any function with finite energy. In fact, if η is such a function then there exists a sequence of functions $\eta_n \in D_A$ such that $|\eta - \eta_n| \to 0$ and $\|\eta - \eta_n\| \to 0$. For the functions η_n identity (2) is valid:

$$[u_0, \eta_n] = (f, \eta_n).$$

Proceeding to the limit as $n \to \infty$ and using the theorem of the continuity of scalar product (theorem 2 § 7) (this theorem and its proof extend directly to the energy product), we find that identity (2) is valid for an arbitrary function η with finite energy. Now in identity (2) it is possible to put $\eta = w_0$, which gives

$$[u_0, w_0] = (f, w_0). \tag{4}$$

We now recall that w_0 gives the minimum of the functional $F(u)$ in the class of functions with finite energy. By writing $F(u)$ in this case in form (1) and repeating the reasoning of § 11 we arrive at the identity

$$[w_0, \eta] = (f, \eta),$$

where η is any function with finite energy. Now putting $\eta = w_0$ and then $\eta \to u_0$, we obtain

$$[w_0, w_0] = (f, w_0), \tag{5}$$

$$[w_0, u_0] = (f, u_0). \tag{6}$$

Subtracting equation (5) from equation (4) and equation (6) from equation (3) we find

$$[w_0 - u_0, w_0] = 0, \qquad [w_0 - u_0, u_0] = 0.$$

Finally the subtraction of these two equations gives $[w_0 - u_0, w - u_0] = 0$ or $|w_0 - u_0|^2 = 0$. Thus $w_0 = u_0$, which was to be proved.

From what has been proved it follows that *the solution of the equation $Au = f$ is equivalent to finding the function which*

realizes the minimum of the functional (1) *in the class of functions with finite energy.* We can easily convince ourselves that for this new variational problem the fundamental results of §§ 12–14 with certain amplifications still hold true. Namely:

(1) Let the sequence of functions $\varphi_n(P)$ be orthonormal and complete in energy. In distinction from § 12, we shall not stipulate that these functions belong to the field of definition of the operator A and we shall restrict ourselves to the requirement that $\varphi_n(P)$ be functions with finite energy. Then the formula (2) § 12

$$u_0(P)=\sum_{n=1}^{\infty}(f,\ \varphi_n)\varphi_n(P),$$

holds where $u_0(P)$ is the solution of the equation $Au=f(P)$;

(2) Let $u_n(P)$, $n=1,\ 2,\ \ldots$ be the minimizing sequence for the functional (1) and let the functions $u_n(P)$ possess finite energy but not necessarily belong to D_A. Then $u_n(P)\overset{E}{\to}u_0(P)$ and since it is assumed that A is a positive-bounded-below operator then $u_n\overset{\text{mean}}{\longrightarrow}u$ simultaneously.

(3) In the application of the Ritz method it is possible to take as the coordinate functions any functions which possess finite energy not necessarily belonging to D_A provided that they satisfy conditions 1) and 2) § 14. The approximate solution of the equation $Au=f$ thus constructed converges in energy to its exact solution. It is necessary in this case to write the Ritz system of equations in the form (8_2) § 14, since the form (8_1) lacks meaning if u does not belong to D_A.

We now pass on to the question of natural boundary conditions which practically is extremely important. By the definition of the operator the functions from its field of definition satisfy the boundary conditions of the corresponding problems. At the same time the examples of § 16 show that functions with finite energy in some cases necessarily satisfy these boundary conditions. In others this is not so.

Boundary conditions which are necessarily satisfied by functions from the field of definition of a given operator and not necessarily so by the functions of finite energy are known as natural for the given operator. Boundary conditions which are necessarily satisfied by functions with finite energy are sometimes called *principal* boundary conditions. (In elasticity theory the principal boundary conditions are usually termed geometric or kinematic and the natural boundary conditions — dynamic.) Thus the examples of § 16 show that for the operator $-\frac{d^2u}{dx^2}(0<x<1)$ the conditions $u(0)=0$ and $u(1)=0$ are principal and the conditions $u'(0)=0$ and $u'(1)=0$ are natural.

From what has been said above it follows that in the application of the Ritz method there is no necessity to subject the coordinate functions to the natural boundary conditions, it suffices that they satisfy the principal conditions. This simple observation plays an important role in the practical application of the Ritz method since it greatly simplifies the choice of the coordinate functions. In this connexion there arises the necessity of stating a simple criterion which can be used to distinguish the natural and principal boundary conditions. We now give such a criterion for a certain class of differential operators which embraces all the cases of reasonable practical interest. Let the operator A have the form

$$Au = \sum_{k=0}^{s} \sum_{\substack{i_1, i_2, \ldots i_k=1 \\ j_1, j_2, \ldots j_k=1}}^{m} \frac{\partial^k}{\partial x_{i_1} \partial x_{i_2} \ldots \partial x_{i_k}} \times$$

$$\times \left(A_{i_1 i_2 \ldots i_k}^{j_1 j_2 \ldots j_k}(P) \frac{\partial^k u}{\partial x_{j_1} \partial x_{j_2} \ldots \partial x_{j_k}} \right), \tag{7}$$

so that its order equals $2s$ and let this operator be positive-bounded-below for a set of functions which satisfy certain boundary conditions. Then the natural boundary conditions

will be those homogeneous ones into which there enter derivatives of u or order s and higher, and the principal ones — those which contain derivatives of u up to the order $s-1$. Thus for the Laplacian operator ($s=1$) the condition $u|_s=0$ is principal and the condition $\left[\frac{\partial u}{\partial n}+\sigma u\right]_s=0$ is natural; for the biharmonic operator ($s=2$)

$$\nabla^4 u=\frac{\partial^4 u}{\partial x^4}+2\frac{\partial^4 u}{\partial x^2\partial y^2}+\frac{\partial^4 u}{\partial y^4},$$

which plays a fundamental role in the theory of bending of thin plates, the conditions for rigid fixing of the edge

$$u|_S=0, \quad \left.\frac{\partial u}{\partial n}\right|_S=0$$

are principal, and of the conditions for free support of an edge (see § 27)

$$u|_S=0, \quad \left[\nabla^2 u-\frac{1-\sigma}{\varrho}\frac{\partial u}{\partial n}\right]_S=0$$

(ϱ is the radius of curvature of the edge) the first is principal and the second natural.

The equation $Au=f$ can also symbolize a system of equations. In this case in formula (7) we understand u to be a vector function, and $A_{i_1 i_2 \cdots i_k}^{j_1 j_2 \cdots j_k}$ a matrix. The test above for distinguishing principal conditions from natural still holds in this case. The proof of the test stated above lies beyond the scope of this book. (For further details see the author's book [11] § 35.)

It is possible to indicate a procedure allowing us to verify in each concrete case whether the given boundary conditions are natural or not. This procedure consists in the following. The expression $(Au, v)=[u, v]$ in which u and v are functions from D_A and satisfy, consequently, the complete boundary conditions, is transformed to a form symmetric in u and v by means of integration by parts and the use of the boundary conditions. Such a transformation is possible since the operator A is symmetric and $(Au, v)=(Av, u)$. Putting $v=u$ in the expression obtained, we obtain a certain expression for the magnitude of $[u, u]$. We suppose that, as often happens, it is also

meaningful and still positive definite for functions which do not satisfy the boundary condition of interest to us. We substitute the expression mentioned for $[u, u]$ in $F(u)$ and suppose that there exists a function u_0 which gives the minimum of this functional in the class of functions which, generally speaking, do not satisfy the given boundary condition. By the usual methods of the variational calculus (see e.g. V.I. Smirnov [2] Ch. II). it is possible to find the necessary conditions which the function u_0 must satisfy. If among them occurs the given boundary condition then this is a natural boundary condition.

Precisely by this method we proved in § 11 that the boundary conditions $u'(0)=u'(1)=0$ are natural for the operator $-\frac{d^2u}{dx^2}$, $0<x<1$.

We now consider a further example. We shall prove that for the equation $-\nabla^2 u=f(P)$ the boundary condition of the mixed problem

$$\left[\frac{\partial u}{\partial n}+\sigma u\right]_S=0, \quad \sigma>0 \tag{8}$$

is natural. Let the functions u and v satisfy condition (8). Then

$$(-\nabla^2 u, v)=-\int_\Omega v\nabla^2 u\,d\Omega,$$

and by Green's formula

$$(-\nabla^2 u, v)=\int_\Omega \nabla u\cdot\nabla v\,d\Omega-\int_S v\frac{\partial u}{\partial n}\,dS.$$

By virtue of condition (8), $\left.\frac{\partial u}{\partial n}\right|_S=-\sigma u|_S$ and so

$$(-\nabla^2 u, v)=\int_\Omega \nabla u\cdot\nabla v\,d\Omega+\int_S \sigma uv\,dS.$$

The right-hand side of this last equation is symmetric in u and v.
Now

$$[u, u]=\int_\Omega (\nabla u)^2 d\Omega+\int_S \sigma u^2 dS,$$

the latter expression being meaningful and positive for any function continuously differentiable in Ω independently of whether it satisfies condition (8) or not.

The functional $F(u)$ can now be represented in the form

$$F(u)=[u, u]-2(f, u)=\int_\Omega [(\nabla u)^2-2fu]d\Omega+\int_S \sigma u^2 dS.$$

We shall consider this functional for the set of all functions which are continuously differentiable twice in Ω, not necessarily satisfying any boundary conditions. Let the function $u_0(P)$ give the minimum of $F(u)$ for the set indicated. We denote by t an arbitrary real number and by $\eta(P)$

an arbitrary function which is twice continuously differentiable in $\overline{\Omega}$. Obviously, $F(u_0+t\eta)\geq F(u_0)$. If the function η is fixed then $F(u_0+t\eta)$ is a function of the independent variable t. The last inequality implies that this function has a minimum at $t=0$. It must follow, then that

$$\frac{\mathrm{d}}{\mathrm{d}t} F(u_0+t\eta)|_{t=0}=0.$$

We easily find

$$F(u_0+t\eta)=F(u_0)+2t\left\{\int_\Omega (\nabla u_0\cdot\nabla\eta-f\eta)\,\mathrm{d}\Omega+\int_S \sigma u_0\eta\,\mathrm{d}S\right\}$$
$$+t^2\left\{\int_\Omega (\nabla\eta)^2\mathrm{d}\Omega+\int_S \sigma\eta^2\mathrm{d}S.\right\}.$$

Differentiating with respect to t, putting $t=0$ and equating the result obtained to zero, we find

$$\int_\Omega (\nabla u_0\cdot\nabla\eta-f\eta)\mathrm{d}\Omega+\int_S \sigma u_0\eta\,\mathrm{d}S=0. \tag{9}$$

By Green's formula

$$\int_\Omega \nabla u_0\cdot\nabla\eta\,\mathrm{d}\Omega=-\int_\Omega \eta\nabla^2 u_0\mathrm{d}\Omega+\int_S \eta\frac{\partial u_0}{\partial n}\,\mathrm{d}S.$$

Equation (9) assumes the form

$$-\int_\Omega \eta(\nabla^2 u_0+f)\mathrm{d}\Omega+\int_S \eta\left(\frac{\partial u_0}{\partial n}+\sigma u_0\right)\mathrm{d}S=0.$$

Because $\eta(P)$ is arbitrary, it follows that $\nabla^2 u_0+f=0$ in Ω and $\dfrac{\partial u_0}{\partial n}+\sigma u_0=0$ in S which was to be proved.

§ 18. Non-homogeneous boundary conditions

Up to now in discussing the equation $Au=f$ we have assumed that the operator A is defined for a set of functions which satisfy homogeneous boundary conditions. This has obliged us to consider only problems in which the conditions at the boundary of the region are homogeneous. In practice, however, non-homogeneous boundary conditions play an important role and it is therefore of interest to explain how

the energy method can be extended for the case of non-homogeneous boundary conditions. We consider the equation

$$Lu=f(P). \tag{1}$$

For illustration we suppose that L is a differential operator of some order k; as its field of definition D_L we take the totality of functions having continuous derivatives of order 1, 2, ..., k--1 and piecewise continuous derivative of order k in the closed domain $\overline{\Omega}=\Omega+S$. We do not subject the functions entering into D_L to any boundary conditions. We shall assume that the operator L is linear. We suppose that equation (1) must be integrated under the boundary conditions

$$G_1u|_S=g_1, \quad G_2u|_S=g_2, \ \ldots, \ G_ru|_S=g_r, \tag{2}$$

where $G_1, G_2, \ldots, G_r$ are linear operators, and $g_1, g_2, \ldots, g_r$ are given functions. The number of the boundary conditions is determined by the order of equation (1) and also by whether the required function $u(P)$ is scalar or vectorial.

We shall solve our problem, making the following extremely important assumption: *there exists a function* $\psi(P)$ *which, together with its derivatives of order up to* $k-1$ *inclusive is continuous in* $\overline{\Omega}$ *which has its derivatives of order* k *piecewise continuous in* $\overline{\Omega}$, *and which satisfies the boundary conditions of the problem, so that*

$$G_1\psi|_S=g_1, \quad G_2\psi|_S=g_2, \ \ldots, \quad G_r\psi|_S=g_r. \tag{3}$$

We put $u-\psi=v$. The new unknown function $v(P)$ satisfies the differential equation

$$Lv=f_1(P); \quad f_1(P)=f(P)-L\psi \tag{4}$$

and the homogeneous boundary conditions

$$G_1v|_S=0, \quad G_2v|_S=0, \quad \ldots, \quad G_rv|_S=0. \tag{5}$$

We assume that *the operator* L *is positive definite for the set of functions which satisfy conditions* (5). According to the

minimal functional theorem the solution of equation (4) under boundary conditions (5) is equivalent to finding the function in the set which minimizes the functional

$$F(v)=(Lv,\ v)-2(v,\ f_1).$$

We now replace v by $u-\psi$, and f_1 by $f-L\psi$:

$$F(v)=(Lu-L\psi,\ u-\psi)-2(u-\psi,\ f-L\psi)$$
$$=(Lu,\ u)-2(u,\ f)+(u,\ L\psi)-(Lu,\ \psi)+2(\psi,\ f)-(L\psi,\ \psi).$$

We can usually succeed in establishing a formula for the operator L, analogous to Green's formula (see § 2) by virtue of which the expression

$$(u,\ L\psi)-(Lu,\ \psi)=\int_{\Omega}(uL\psi-\psi Lu)\mathrm{d}\Omega$$

can be transformed into some surface integral which we denote by

$$\int_S R(u,\ \psi)\mathrm{d}S.$$

The expression $R(u,\ \psi)$ depends upon the form of the operator L. Using boundary conditions (2) and (3), it is often possible to represent $R\ (u,\ \psi)$ in the form

$$R(u,\ \psi)=N(u)+M.$$

$N(u)$ depends only upon u and upon the functions g_1, $g_2,\ \ldots,\ g_r$ which enter into boundary conditions (2). M does not depend upon u but it can depend upon ψ. In such a case $F(v)$ is reduced to the form

$$F(v)=(Lu,\ u)-2(u,\ f)$$
$$+\int_S N(u)\mathrm{d}S+\left[2(\psi,\ f)-(L\psi,\ \psi)+\int_S M\,\mathrm{d}S\right].$$

The expression in square brackets is a constant quantity (although perhaps unknown in practice if we cannot find the function ψ). Thus the minimizing of the functional $F(v)$ is

equivalent to the minimizing of another functional which we denote by $\Phi(u)$:

$$\Phi(u)=(Lu,\ u)-2(u,\ f)+\int_S N(u)\mathrm{d}S$$

$$=\int_\Omega (uLu-2uf)\mathrm{d}\Omega+\int_S N(u)\mathrm{d}S. \tag{6}$$

The minimum is with respect to functions which satisfy boundary conditions (2).

NOTE 1: Functional (6) can be constructed without knowing the function. However, in order that the problem of the minimum of this functional has meaning it is necessary to be certain that such a function exists, otherwise the variational problem can have no solution as Hadamard's problems (see introduction) show. In a number of cases it has been proved that the function ψ exists if the surface S is sufficiently smooth and the functions $g_1, g_2, \ldots, g_r$ which enter into condition (2) are differentiable a sufficient number of times in any direction tangential to the surface S. We shall not dwell further on this question.

NOTE 2: It may happen that some of conditions (5) arise naturally for the functional $F(v)$. The non-homogeneous conditions (2) corresponding to them will then be natural for the functional $\Phi(u)$, the minimum of which can therefore be sought in the class of functions which satisfy all conditions (2) besides perhaps the natural ones. This is a wider class than that indicated above.

As an example we pose the problem of integrating the Laplace equation

$$-\nabla^2 u=0 \tag{7}$$

under the boundary conditions

$$u|_S=g(P). \tag{8}$$

As is known, this problem is termed the Dirichlet problem. We have put a minus sign on the left of equation (7) since under the boundary condition $u|_S=0$ the operator $-\nabla^2$ is

positive definite. In our case $f \equiv 0$, and the second term in (6) vanishes. In order to find $N(u)$ we construct the expression

$$\int_{\Omega} (uL\psi - \psi Lu)\, d\Omega = \int_{\Omega} (\psi \nabla^2 u - u \nabla^2 \psi) d\Omega,$$

because in our problem $L = -\nabla^2$. By Green's formula (formula (10) § 2)

$$\int_{\Omega} (\psi \nabla^2 u - u \nabla^2 \psi) d\Omega = \int_{S} \left(\psi \frac{\partial u}{\partial n} - u \frac{\partial \psi}{\partial n} \right) dS,$$

where n is the external normal to S. Thus in our case

$$R(u, \psi) = \psi \frac{\partial u}{\partial n} - u \frac{\partial \psi}{\partial n}.$$

By virtue of condition (8) $u|_S = g(P)$; the function such that $\psi|_S = g(P)$ also satisfies the same boundary condition. This gives $R(u, \psi) = g \frac{\partial u}{\partial n} - g \frac{\partial \psi}{\partial n}$. Hence it is clear that in our case $N(u) = g \frac{\partial u}{\partial n}$ and $M = g \frac{\partial \psi}{\partial n}$. This gives

$$\Phi(u) = -\int_{\Omega} u \nabla^2 u d\Omega + \int_{S} g \frac{\partial u}{\partial n} dS. \tag{9}$$

Expression (9) can be simplified if we use formula (9) § 2:

$$-\int_{\Omega} u \nabla^2 u d\Omega = \int_{\Omega} (\triangle u)^2 d\Omega - \int_{S} u \frac{\partial u}{\partial n} dS. \tag{10}$$

We substitute this in (9). Noting that $u|_S = g(P)$ we find that the surface integral vanishes. The Dirichlet problem thus reduces to finding the minimum of the integral

$$\Phi(u) = \int_{\Omega} (\triangle u)^2 d\Omega \tag{11}$$

with respect to functions which at the boundary of the region S assume the given value $g(P)$. Historically the Dirichlet

problem was the first to which the energy method was applied.

We now consider the further problem of integrating expression (7) under the boundary conditions

$$\left.\frac{\partial u}{\partial n}\right|_S = h(P) \tag{12}$$

(the Neumann problem). As before $f=0$ and $R(u, \psi) = \psi \dfrac{\partial u}{\partial n} - u\dfrac{\partial \psi}{\partial n}$ but this time the boundary equation gives $\left.\dfrac{\partial u}{\partial n}\right|_S = \left.\dfrac{\partial \psi}{\partial n}\right|_S = h$ and $R(u, \psi) = \psi h - uh$. Hence $N(u) = -uh$, $M = \psi h$ and by formula (6):

$$\Phi(u) = -\int_\Omega u \nabla^2 u \, d\Omega - \int_S uh \, dS.$$

We transform the volume integral according to formula (10). Using condition (12) again we find finally

$$\Phi(u) = \int_\Omega (\nabla u)^2 d\Omega - 2\int_S uh \, dS. \tag{13}$$

The boundary condition (12) is natural for the functional (13) and its minimum can be sought in the class of functions which do not satisfy the boundary conditions.

§ 19. The existence of a solution of the variational problem

In all the foregoing we have started from the assumption that the equation considered

$$Au = f$$

has a solution possessing finite energy. Hence it followed that the problem of the minimum of the functional

$$F(u) = (Au, u) - 2(u, f) = [u, u] - 2(u, f),$$

has a solution, if the operator A is positive definite. In the present section we shall show that *the problem of the minimum*

of the functional $F(u)$ has a solution in the class of functions with finite energy only if the operator A is positive-bounded-below.

We represent the functional $F(u)$ in the form

$$F(u)=\|u\|^2-2(u, f). \tag{1}$$

We shall prove, first of all, that this functional is bounded below. By Cauchy's inequality (formula (8) § 3) $|(u, f)| \leq \|u\| \|f\|$ but the operator A is positive-bounded-below and so $\|u\| \leq \|u\|/\gamma$ (γ is the constant in inequality (7) § 6), and

$$F(u) \geq \|u\|^2-\frac{2}{\gamma}\|u\|\|f\| = \left(\|u\|-\frac{1}{\gamma}\|f\|\right)^2-\frac{\|f\|^2}{\gamma^2}.$$

By removing the first term on the right we strengthen the inequality. Thus

$$F(u) \geq -\frac{\|f\|^2}{\gamma^2}.$$

For any $u(P)$ from the class of functions with finite energy, the functional $F(u)$ is not less than the constant magnitude $-\frac{\|f\|^2}{\gamma^2}$. This means also that the functional $F(u)$ is bounded below. Hence it follows that this functional has an exact lower bound (cf. 1, **42**) which we shall denote by d.

As was proved in § 17 this bound is unaltered if we consider the value of $F(u)$ not with respect to functions with finite energy but for the class D_A of functions constituting the field of definition of the operator A.

By the definition of exact lower bound, for any positive number ε it is possible to find a function with finite energy which we denote by $u(P, \varepsilon)$ such that $F(u(P, \varepsilon)) < d+\varepsilon$. At the same time, obviously $F(u(P, \varepsilon)) \geq d$. We put $\varepsilon=\frac{1}{n}$, where n is any positive integer; the corresponding function $u\left(P, \frac{1}{n}\right)$ we write more briefly $u_n(P)$. Then

$$d \leq F(u_n) < d+\frac{1}{n}, \qquad n=1, 2, \ldots \tag{2}$$

Proceeding to the limit as $n \to \infty$ we find that

$$\lim_{n\to\infty} F(u_n) = d. \tag{3}$$

Equation (3) means that *the functions $u_n(P)$ constitute a minimizing sequence for the functional $F(u)$.*

The next step consists in the proof of the fact that our minimizing sequence converges in energy to some limit. We specify an arbitrary number $\varepsilon > 0$ and choose an integer N such that $\frac{1}{N} < \frac{\varepsilon}{2}$. Then if $n > N$ and $m > M$, $\frac{1}{n} < \frac{\varepsilon}{2}$, $\frac{1}{m} < \frac{\varepsilon}{2}$ and consequently

$$d \leq F(u_n) < d + \frac{\varepsilon}{2}, \qquad d \leq F(u_m) < d + \frac{\varepsilon}{2}. \tag{4}$$

Let t denote an arbitrary number between zero and unity. We consider the function $w_t(P) = t\,u_n(P) + (1-t)u_m(P)$ and we construct the expression

$$F(w_t) = [tu_n + (1-t)u_m,\ tu_n + (1-t)u_m] - 2(tu_n + (1-t)u_m,\ f).$$

Removing the brackets we easily find

$$F(w_t) = \|u_n - u_m\|^2 t^2 + 2\{[u_n - u_m,\ u_m] - (u_n - u_m,\ f)\}t + \|u_m\|^2 - 2(u_n,\ f). \tag{5}$$

Thus $F(w_t)$ is a quadratic trinominal in t with a positive coefficient for t^2. The graph of such a trinominal is a parabola, concave upwards. If any chord is drawn on such a parabola (Fig. 7) then a chord shorter than this lies below it. Hence there emerges the following conclusion: We consider a chord, the abscissae of the ends of which are $t=0$, $t=1$. For any t, $0 < t < 1$, $F(w_t)$ is less than the largest of the quantities $F(w_0)$ and $F(w_1)$. But $w_0 = u_m$, $w_1 = u_n$, in virtue of inequalities (4) each of the quantities $F(u_n)$ and $F(u_m)$ is less than $d + \frac{\varepsilon}{2}$. Hence it follows that $F(w_t) < d + \frac{\varepsilon}{2}$, $0 < t < 1$. At the

same time $F(w_t) \geq \inf F(u) = d$. In particular we put $t = \frac{1}{2}$. Then $w_{\frac{1}{2}} = \frac{u_m + u_n}{2}$ and

$$d \leq F\left(\frac{u_m + u_n}{2}\right) < d + \frac{\varepsilon}{2}. \tag{6}$$

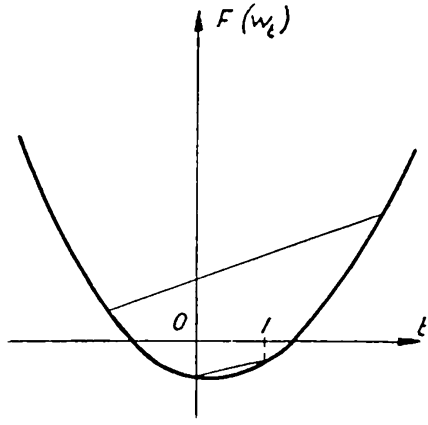

FIG. 7

From inequalities (4) and (6) it easily follows that

$$-\varepsilon < F(u_n) + F(u_m) - 2F\left(\frac{u_m + u_n}{2}\right) < \varepsilon, \quad m, n > N,$$

and this inequality means that

$$\lim_{n, m \to \infty} \left[F(u_n) + F(u_m) - 2F\left(\frac{u_m + u_n}{2}\right)\right] = 0. \tag{7}$$

We now evaluate the quantity under the limit sign in (7). By formula (1)

$$F(u_n) + F(u_m) - 2F\left(\frac{u_m + u_n}{2}\right) = \|u_n\|^2 + \|u_m\|^2 - \frac{1}{2}\|u_m + u_n\|^2. \tag{8}$$

Further

$$\|u_n + u_m\|^2 = [u_n + u_m, u_n + u_m] = \|u_n\|^2 + 2[u_n, u_m] + \|u_m\|^2.$$

Substituting this in (8) we obtain

$$F(u_n)+F(u_m)-2F\left(\frac{u_n+u_m}{2}\right)=\frac{1}{2}\|u_n-u_m\|^2.$$

Now from the inequality (7) it follows that

$$\lim_{n,\,m\to\infty}\|u_n-u_m\|=0. \tag{9}$$

By virtue of what was proved in § 16 there exists a function $u_0(P)$ with finite energy such that $u_n \xrightarrow{E} u_0$ and $u_n \xrightarrow{\text{mean}} u_0$.

We now prove that the function $u_0(P)$ minimizes the functional $F(u)$. Since $u_n \xrightarrow{E} u_0$, $\|u_n\|\to\|u_0\|$. At the same time by the continuity of scalar multiplication (§ 3), we shall have $(u_n, f)\to(u_0, f)$.

Now

$$F(u_n)=\|u_n\|^2-2(u_n, f)\to\|u_0\|^2-2(u_0, f)=F(u_0)$$

and from (3) it follows that

$$F(u_0)=d=\inf F(u),$$

which was to be proved.

Thus in the class of functions with finite energy there exists a function $u_0(P)$ which minimizes functional (1). If this function also belongs to the field of definition D_A of the operator then by virtue of the minimum functional theorem (§ 11) this function also satisfies the equation $Au=f$. In the general case the function $u_0(P)$ should be considered as the *generalized solution* of this equation. The methods presented in the foregoing paragraphs of the present chapter are entirely applicable when the equation $Au=f$ has a generalized solution and not an ordinary solution.

CHAPTER IV

MAJOR APPLICATIONS OF THE ENERGY METHOD

§ 20. Boundary value problems for an ordinary differential equation

Many problems in mathematical physics reduce to finding the solution of an ordinary differential equation, which satisfies certain boundary conditions which can always be made homogeneous. We can apply the methods expounded in the previous paragraph to our problems if we can establish that, under certain restrictions, the operator associated with the problems considered which is defined by the left-hand side of the differential equation and by the boundary conditions, is symmetric and positive-bounded-below. We shall go into this in the present paragraph.

We consider the ordinary differential equation

$$Lu = -\frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}u}{\mathrm{d}x}\right) + r(x)u = f(x). \tag{1}$$

We shall seek a solution which together with its first derivative is continuous over the range $a \leq x \leq b$ and which satisfies at the ends of the range the boundary conditions

$$\left.\begin{aligned} \alpha u'(a) - \beta u(a) &= 0, \\ \gamma u'(b) + \delta u(b) &= 0, \end{aligned}\right\} \tag{2}$$

where α, β, γ, δ are constants.

We make further the following assumptions:

(a) $p(x)$, $p'(x)$ and $r(x)$ are continuous over the range $a \leq x \leq b$;

(b) within the range $p(x) \geq 0$, $r(x) \geq 0$;

(c) the function $p(x)$ can vanish at a number of points in the range $a \leq x \leq b$ provided however that the integral

$$A = \int_a^b \frac{dx}{p(x)} \tag{3}$$

converges (and consequently has a finite value). In particular, this requirement is always satisfied if $p(x)$ vanishes nowhere in the segment $a \leq x \leq b$;

(d) the constants α, β, γ, δ are non-negative, and are such that neither of the products $\alpha\beta$, $\gamma\delta$ vanishes.

For the field of definition D_A of the operator A we take the set of functions which together with their first and second derivatives are continuous over the range $a \leq x \leq b$, and which satisfy boundary conditions (2). We shall prove that the operator L is symmetric. To this end we form the scalar product (Lu, v) where both the functions u and $v \in D_L$. In particular both satisfy conditions (2). We have

$$(Lu, v) = -\int_a^b v \frac{d}{dx}\left(p(x)\frac{du}{dx}\right)dx + \int_a^b r(x)u(x)v(x)dx.$$

Integrating the first integral by parts and using conditions (2)

$$(Lu, v) = \int_a^b \left[p(x)\frac{du}{dx}\frac{dv}{dx} + r(x)uv\right]dx + \\ + \frac{\beta}{\alpha}p(a)u(a)v(a) + \frac{\delta}{\gamma}p(b)u(b)v(b). \tag{4}$$

We assume here that $\alpha \neq 0$, $\gamma \neq 0$. If $\alpha = 0$ or $\gamma = 0$ then $u(a) = 0$ or $u(b) = 0$ and the corresponding terms on the right in (4) can be omitted. In particular, if $\alpha = \gamma = 0$ then conditions (2) assume the form

$$u(a) = u(b) = 0 \tag{2$_1$}$$

and

$$(Lu, v) = \int_a^b \left[p(x) \frac{du}{dx} \frac{dv}{dx} + ruv \right] dx. \tag{4_1}$$

The expression on the right in (4) is symmetric in u and v; hence it is easily concluded that $(Lu, v) = (u, Lv)$, i.e. that the operator L is symmetric.

Putting $u(x) = v(x)$ in (4), we find that

$$(Lu, u) = \int_a^b \left[p(x) \left(\frac{du}{dx} \right)^2 + r(x)u^2(x) \right] dx + \frac{\beta}{\alpha} p(a)u^2(a) + \frac{\delta}{\gamma} p(b)u^2(b). \tag{5}$$

We shall prove that if at least one of the numbers β and δ is different from zero and $p(a)$ and $p(b)$ are positive, then the operator L is positive-bounded-below. We consider first of all the very simple special case when the boundary conditions have the form (2_1), and $p(x) \geq p_0$ where p_0 is a positive constant. In this case formula (5) simplifies to:

$$(Lu, u) = \int_a^b \left[p(x) \left(\frac{du}{dx} \right)^2 + r(x)u^2(x) \right] dx. \tag{5_1}$$

In formula (5_1) on the right we drop the non-negative second term and replace $p(x)$ by the constant p_0 which is always less. Equality (5_1) then becomes the inequality

$$(Lu, u) \geq p_0 \int_a^b \left(\frac{du}{dx} \right)^2 dx = p_0 \left\| \frac{du}{dx} \right\|^2. \tag{6}$$

It remains to estimate $\left\| \frac{du}{dx} \right\|$. Since $u(a) = 0$

$$u(x) = \int_a^x u'(t)dt, \quad u'(t) = \frac{du(t)}{dt}.$$

By using Cauchy's inequality (formula (8_1) § 3) over the range $a \leq x \leq b$

$$u^2(x) \leq (x-a)\int_a^x u'^2(t)\,dt \leq (x-a)\int_a^b u'^2(t)\,dt = (x-a)\|u'\|^2.$$

Integrating this between the limits a and b we find that

$$\|u\|^2 \leq \frac{(b-a)^2}{2}\|u'\|^2. \tag{7}$$

Substituting this in (5), we obtain the inequality

$$(Lu,\ u) \geq \gamma^2\|u\|^2; \quad \gamma^2 = \frac{2p_0}{(b-a)^2},$$

which proves that the operator L is positive-bounded-below.

We now turn to the general case. We suppose that $\beta > 0$. In formula (5) on the right we drop the second term under the integral sign and the last term. The right-hand side of (5) is smaller, in general, and we obtain the inequality

$$(Lu,\ u) \geq \int_a^b p(x)\left(\frac{du}{dx}\right)^2 dx + \frac{\beta}{\alpha}p(a)u^2(a)$$

$$\geq B\left[\int_a^b pu'^2\,dx + u^2(a)\right], \tag{8}$$

where B is the lesser of the numbers $\frac{\beta}{\alpha}p(a)$ and 1. We have

$$u(x) = u(a) + \int_a^x u'(t)\,dt$$

and $(a+b)^2 \leq (a+b)^2 + (a-b)^2 = 2(a^2+b^2)$, thus

$$u^2(x) \leq 2u^2(a) + 2\left(\int_a^x u'(t)\,dt\right)^2.$$

Further

$$\left(\int_a^x u'(t)\mathrm{d}t\right)^2 = \left(\int_a^x \frac{1}{\sqrt{(p(t))}}\sqrt{(p(t))}u'(t)\mathrm{d}t\right)^2,$$

and by Cauchy's inequality

$$\left(\int_a^x u'(t)\mathrm{d}t\right)^2 \leq \int_a^x \frac{\mathrm{d}t}{p(t)} \int_a^x p(t)u'^2(t)\mathrm{d}t.$$

Increasing the upper limit of integration on the right only strengthens the inequality. Therefore we replace x by its maximum value b on the right. Thus we obtain

$$\left(\int_a^x u'(t)\mathrm{d}t\right)^2 \leq \int_a^b \frac{\mathrm{d}t}{p(t)} \int_a^b p(t)u'^2(t)\mathrm{d}t = A\int_a^b p(t)u'^2(t)\mathrm{d}t$$

and consequently

$$u^2(x) \leq 2A\int_a^b p(t)u'^2(t)\mathrm{d}t + 2u^2(a).$$

Integrating this with respect to x in the range from a to b, we find

$$\|u\|^2 \leq C\left[\int_a^b p(t)u'^2(t)\mathrm{d}t + u^2(a)\right], \tag{9}$$

where C is the larger of the two numbers $2A(b-a)$ and $2(b-a)$. The comparison of inequalities (8) and (9) proves that

$$(Lu, u) \geq \frac{B}{C}\|u\|^2, \tag{10}$$

i.e. that the operator L is positive-bounded-below.

It was assumed above that $\alpha \neq 0$ and $\gamma \neq 0$. If one of the numbers α and γ (or both of these numbers) vanishes, then the statement about the operator L being positive-bounded-

below is still true, the reasoning becoming even simpler. Let, for example, $\alpha=0$.

The first bounof dary conditions (2) then assumes the form $u(a)=0$. In place of (8) and (9) we obtain the simple inequality

$$(Lu, u)\geq\int_a^b p(x)u'^2(x)\mathrm{d}x, \quad \|u\|^2\leq A(b-a)\int_a^b p(x)u'^2(x)\mathrm{d}x.$$

Hence inequality (10) follows in which we need only replace $\frac{B}{C}$ by $\frac{1}{A(b-a)}$.

In virtue of the minimum functional theorem (§ 11) the problem of integrating equation (1) under boundary conditions (2) reduces if $\alpha\neq 0$ and $\gamma\neq 0$, to the finding of the function which minimizes the functional:

$$F(u)=(Lu, u)-2(f, u)=\frac{\beta}{\alpha}p(a)u^2(a)+\frac{\delta}{\gamma}p(b)u^2(b)$$
$$+\int_a^b[p(x)u'^2(x)+r(x)u^2(x)-2f(x)u(x)]\mathrm{d}x. \qquad (11)$$

If $\alpha\neq 0$ or $\gamma\neq 0$, then the corresponding boundary condition (2), is natural as was stated in § 12. If $\alpha=0$ or $\gamma=0$ then the corresponding boundary condition is principal. By virtue of the results of § 19 our boundary problem has a solution. It can be constructed, approximately, or exactly, by one of the methods given in §§ 12–15.

Let $u_0(x)$ be the exact solution of the problem, and let $u_n(x)$ be its approximate solution constructed as just indicated. Then, as we know, $u_n\xrightarrow{\mathrm{E}}u$. We now consider the nature of energy convergence in the given case. We know that $u_0(x)$ is a function with finite energy and consequently with a finite norm $\left(\text{since } \|u\|\leq\frac{1}{\gamma}|u|\right)$ but none of its differential properties are known to us. As regards the approximate solutions $u_n(x)$ they can always be constructed so as to

belong to the field of definition of the operator. In such a case the functions $u_n(x)$ have continuous first (and even second) derivatives.

We restrict ourselves to the case where the inequality $p(x) \geq p_0$ is satisfied in the interval $a \leq x \leq b$, p_0 being a positive constant.

By formula (5) we have

$$\|u\|^2 = \int_a^b [p(x)u'^2(x) + r(x)u^2(x)]\mathrm{d}x + \frac{\beta}{\alpha}u^2(a) + \frac{\delta}{\gamma}u^2(b).$$

As was stated $u_n \xrightarrow{E} u_0$, or, what is the same thing,

$$\lim_{n\to\infty} \|u_n - u_0\| = 0.$$

By the triangle inequality (§ 8)

$$\|u_n - u_m\| \leq \|u_n - u_0\| + \|u_m - u_0\|$$

and consequently

$$\lim_{n,\, m \to \infty} \|u_n - u_m\|^2 = 0.$$

Using formula (11) this can be given the form

$$\lim_{n,\, m \to \infty} \left\{ \int_a^b [p(x)(u_n'(x) - u_m'(x))^2 + r(x)(u_n(x) - u_m(x))^2]\mathrm{d}x \right.$$
$$\left. + \frac{\beta}{\alpha} p(a)[u_n(a) - u_m(a)]^2 + \frac{\delta}{\gamma} p(b)[u_n(b) - u_m(b)]^2 \right\} = 0.$$

The terms on the left are non-negative and thus each of them separately tends to zero. In particular

$$\lim_{n,\, m \to \infty} \int_a^b p(x)[u_n'(x) - u_m'(x)]^2\mathrm{d}x = 0; \quad \lim[u_n(a) - u_m(a)]^2 = 0. \quad (12)$$

We have obviously

$$\int_a^b [u_n'(x) - u_m'(x)]^2\mathrm{d}x \leq \frac{1}{p_0} \int_a^b p(x)[u_n'(x) - u_m'(x)]^2\mathrm{d}x;$$

whence

$$\lim_{n, m \to \infty} \int_a^b [u_n'(x) - u_m'(x)]^2 dx = 0. \tag{13}$$

Thus in the case considered, there follows from the energy convergence of the functions $u_n(x)$ the convergence in the mean of their first derivatives. If the function $f(x)$ has a finite norm then it can be shown that $u_0(x)$ has a continuous first derivative $u_0'(x)$ and that $u_n'(x) \xrightarrow{\text{mean}} u_0'(x)$.

We now consider the difference

$$u_n(x) - u_m(x) = u_n(a) - u_m(a) + \int_a^x [u_n'(t) - u_m'(t)] dt.$$

In virtue of the inequality $(a+b)^2 \leq 2(a^2+b^2)$ and Cauchy's inequality

$$[u_n(x) - u_m(x)]^2 \leq 2[u_n(a) - u_m(a)]^2 + 2\left\{\int_a^x [u_n'(t) - u_m'(t) dt\right\}^2$$

$$\leq 2[u_n(a) - u_m(a)]^2 + 2(x-a) \int_a^x [u_n'(t) - u_m'(t)]^2 dt.$$

By replacing x by b we strengthen the inequality to

$$[u_n(x) - u_m(x)]^2 \leq 2[u_n(a) - u_m(a)]^2 + 2(b-a) \int_a^b [u_n'(t) - u_m'(t)]^2 dt. \tag{14}$$

Hence it is clear that the sequence $u_n(x)$, $n = 1, 2, \ldots$ converges uniformly to some limiting function $\bar{u}(x)$, since the right-hand side of inequality (14) does not depend upon x and tends to zero as $n \to \infty$.

It is easily seen that $\bar{u}(x) = u_0(x)$. In fact, since $u_n(x) \xrightarrow{E} u_0(x)$ then simultaneously $u_n(x) \xrightarrow{\text{mean}} u_0(x)$.

On the other hand from uniform convergence there also follows convergence in the mean (§ 7), so that $u_n(x) \xrightarrow{\text{mean}} u(x)$,

and since one and the same sequence cannot tend in the mean to two different limits, $\bar{u}(x) \equiv u_0(x)$.

Thus, if $p(x) \geq p_0 > 0$, energy convergence implies the uniform convergence of the functions themselves and convergence in the mean of their derivatives.

NOTE: If we reject the stipulation that $p(x) \geq p_0 > 0$, but subject $p(x)$ to the condition that the integral (3) converges, then, as before, it can be proved that $u_n(x) \to u_0(x)$ uniformly, but instead of equation (13) we can deduce only the weaker equation (12). From this in turn we can obtain that

$$\lim_{n \to \infty} \int_a^b p(x)[u_n'(x) - u_0'(x)]^2 \mathrm{d}x = 0.$$

We pass on to an equation of arbitrary even order

$$Lu = \sum_{k=0}^{m} (-1)^k \frac{\mathrm{d}^k}{\mathrm{d}x^k} \left[p_k(x) \frac{\mathrm{d}^k u}{\mathrm{d}x^k} \right] = f(x), \tag{15}$$

and confine ourselves to the simplest of boundary conditions

$$\begin{aligned} u(a) = u'(a) = \ldots = u^{(m-1)}(a) \\ = u(b) = u'(b) = \ldots = u_s^{(m-1)}(b) = 0. \end{aligned} \tag{16}$$

We shall consider that the coefficients $p_k(x)$, $k = 0, 1, 2, \ldots, m-1$ are non-negative and that $p_m(x)$ is strictly positive; further we shall assume that $f(x)$ has a finite norm. Integrating by parts and using condition (16), we find that

$$\begin{aligned} (Lu, u) &= \sum_{k=0}^{m} \int_a^b p_k(x) \left(\frac{\mathrm{d}^k u}{\mathrm{d}x^k} \right)^2 \mathrm{d}x \\ &\geq \int_a^b p_m(x) \left(\frac{\mathrm{d}^m u}{\mathrm{d}x^m} \right)^2 \mathrm{d}x \geq p_0 \| u^{(m)} \|^2, \end{aligned} \tag{17}$$

where this time

$$p_0 = \min p_m(x);$$

we assume that $p_0 > 0$.

Inequality (7) is valid for any function equal to zero at

$x=a$. The functions $u(x)$, $u'(x)$, ..., $u^{m-1}(x)$ satisfy this requirement and thus there occur the inequalities

$$\|u\| \leq \frac{b-a}{\gamma(2)} \|u'\|, \quad \|u'\| \leq \frac{b-a}{\gamma(2)} \|u''\|, \ \ldots, \ \|u^{(m-1)}\| \leq \frac{b-a}{\gamma(2)} \|u^{(m)}\|,$$

whence it readily follows that $\|u^{(m)}\| \geq \left(\frac{\gamma(2)}{b-a}\right)^m \|u\|$. Substituting this in (17), we obtain

$$(Lu, u) \geq \gamma^2 \|u\|^2, \quad \gamma = \gamma(p_0) \left(\frac{\gamma(2)}{b-a}\right)^m.$$

The latter inequality proves that under conditions (16) the operator (15) is positive-bounded-below hence, as we have often recalled, the existence of a solution and the convergence of the minimizing sequence follow. We now explain the nature of this convergence.

Inequality (17) can be given the form

$$|u| \geq \gamma(p_0) \|u^{(m)}\|. \tag{18}$$

Let $u_n(x)$, $n=1, 2, \ldots$ be the minimizing sequence. By theorem 1 § 13 $|u_n - u_0| \to 0$ where $u_0(x)$ is the exact solution of the problem. By the triangle inequality (§ 3) (which is also true for the energy norm) we have

$$|u_n - u_k| \leq |u_n - u_0| + |u_k - u_0|$$

and consequently

$$\lim_{k, n \to \infty} |u_n - u_k| = 0. \tag{19}$$

In virtue of conditions (16) $u^{(m-1)}(a) = u_k^{(m-1)}(a) = 0$, hence

$$u_n^{(m-1)}(x) - u_k^{(m-1)}(x) = \int_a^x [u_n^{(m)}(t) - u_k^{(m)}(t)] \mathrm{d}t$$

and by Cauchy's inequality

$$|u_n^{(m-1)}(x) - u_k^{(m-1)}(x)|^2 \leq (x-a) \int_a^x [u_n^{(m)}(t) - u_k^{(m)}(t)]^2 \mathrm{d}t$$

$$\leq (b-a) \int_a^b [u_n^{(m)}(t) - u_k^{(m)}(t)]^2 \mathrm{d}t = (b-a) \|u_n^{(m)} - u_k^{(m)}\|^2.$$

Now from (18) and (19) it follows that the $(m-1)$th derivatives of $u_n(x)$ converge uniformly. But then by Weierstras's integration theorem for uniformly converging sequences (cf. 1, **145**) both the approximate solutions $u_n(x)$ themselves and their derivatives up to order $m-1$ inclusive converge uniformly. From relationships (18) and (19) it also follows that the mth derivatives of $u_n(x)$ converge in the mean.

§ 21. Bending of a beam of variable cross-section, resting on an elastic foundation

The equation for the bending of a beam, resting on an elastic foundation, has the form

$$Lw=\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left[EI(x)\frac{\mathrm{d}^2w}{\mathrm{d}x^2}\right]+Kw=q(x). \tag{1}$$

Here w is the deflexion of the beam at the section with abscissa x, $I(x)$ being the moment of inertia of this section. $I(x)$ will be a constant or a function of x depending upon whether the beam section is constant or variable. For simplicity we assume that nowhere does the beam cross-section degenerate to a point, so that the moment of inertia $I(x)$ never vanishes. E is the Young's modulus of the beam material, K is the yield coefficient for the foundation, and $q(x)$ is the normal loading. We denote the length of the beam by l. If the ends of the beam are rigidly fixed then the following boundary conditions must be satisfied:

$$w(0)=w(l)=0; \qquad w'(0)=w'(l)=0. \tag{2}$$

Equations (1) and (2) are special cases (for $m=2$) of equations (15) and (16) of the previous paragraph. Hence we conclude that the operator defined by the left-hand side of equation (1) and boundary conditions (2) is positive-bounded-below and that the problem of the bending of a beam lying on an elastic foundation and having rigidly fixed ends reduces to the problem of minimizing the functional

$$F(w)=(Lw, w) - 2(w, q)$$

with respect to functions satisfying the boundary conditions (2). If we use formula (17) of the previous paragraph, then $F(w)$ can be represented in the form

$$\overline{F(w)} = \int_a^b [EI(x)w''^2 + Kw^2 - 2q(x)w]\mathrm{d}x. \tag{3}$$

The solution of the problem of finding the minimum of the functional (3) can be obtained by the Ritz method, for which it is necessary to choose a system of coordinate functions $\varphi_n(x)$, complete in energy. Since conditions (2) are principal, the coordinate functions necessarily satisfy them. If, for example, we put

$$\varphi_k(x) = (l-x)^2 x^{k+1}, \qquad k = 1, 2, \ldots, \tag{4}$$

it can be shown that system (4) is complete in energy. We obtain an approximate solution of our problem by putting

$$w_n = \sum_{k=1}^{n} a_k \varphi_k(x) = (l-x)^2 \sum_{k=1}^{n} a_k x^{k+1} \tag{5}$$

and by determining the coefficients of system (8) § 14, which we write thus

$$\sum_{k=1}^{n} a_k A_{ik} = b_i, \quad i = 1, 2, \ldots, n. \tag{6}$$

Here

$$b_k = (q, \varphi_k) = \int_0^l q(x)(l-x)^2 x^{k+1}\mathrm{d}x. \tag{7}$$

The coefficients A_{ik} can be represented in one of two forms, corresponding to systems (8_1) and (8_2) § 14:

$$A_{ik} = (L\varphi_i, \varphi_k) = \int_0^l \varphi_k L\varphi_i \,\mathrm{d}x = \int_0^l \varphi_k \left[\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(EI(x)\frac{\mathrm{d}^2\varphi_i}{\mathrm{d}x^2}\right) + K\varphi_i\right]\mathrm{d}x; \tag{8}$$

$$A_{ik} = [\varphi_i, \varphi_k] = \int_0^l \left(EI(x)\frac{\mathrm{d}^2\varphi_i}{\mathrm{d}x^2}\frac{\mathrm{d}^2\varphi_k}{\mathrm{d}x^2} + K\varphi_i\varphi_k\right)\mathrm{d}x.$$

We denote the exact solution of our problem by $w_0(x)$. From the general considerations given at the end of the last paragraph it follows that $w_n(x) \to w_0(x)$ and $w_n'(x) \to w_0'(x)$ uniformly and $w_n''(x) \xrightarrow{\text{mean}} w_0''(x)$. If one of the ends of the beam is free, then (2) is replaced by the condition of the vanishing of the quantities $\frac{d^2w}{dx^2}$ and $\frac{d}{dx}\left(EI(x)\frac{d^2w}{dx^2}\right)$ at the corresponding end. For example, if the end $x=0$ is rigidly fixed and the end $x=l$ is free, then the boundary conditions are

$$w(0)=0, \quad w'(0)=0, \quad w''(l)=0, \quad \frac{d}{dx}\left[EI(x)\frac{d^2w}{dx^2}\right]_{x=l}=0. \qquad (9)$$

It is easily verified that under these conditions the operator on the left of equation (1) is positive-bounded-below as before. We leave the verification of this and the derivation of the conclusions resulting therefrom concerning the solubility of the problem and the applicability of approximate methods for its solution, to the reader.

The results of the present paragraph still hold in the case when the yield coefficient K of the base is variable.

§ 22. Fundamental boundary value problems for the equations of Poisson and Laplace

In previous sections we have frequently considered boundary value problems for the equations of Poisson and Laplace as illustrations of general methods; in the present section we study these problems more systematically.

Consider the Poisson equation

$$-\nabla^2 u = f(P), \qquad (1)$$

where P is a point in some finite m-dimensional region (m can be taken as 2 or 3). We pose for this equation the Dirichlet problem: to find the solution of equation (1),

continuous in the closed domain $\bar{\Omega}=\Omega+S$ which satisfies at the boundary S of this region the boundary condition:

$$u|_S=0. \tag{2}$$

As the field of definition of the Laplacian operator, we take the linear set (denoted by the letter M) of functions satisfying the following conditions: 1) they are continuous together with their first and second derivatives in the closed domain $\bar{\Omega}=\Omega+S$; 2) they vanish on S. It is easily seen that for the linear set M the operator $-\nabla^2$ is positive definite. In fact, by Green's formula (§ 2)

$$(-\nabla^2 u,\ u)=-\int_{\Omega} u\nabla^2 u\,d\Omega=-\int_{S} u\frac{\partial u}{\partial n}\,dS+\int_{\Omega}(\nabla u)^2 d\Omega, \tag{3}$$

where n is the external normal to S. The surface integral disappears, since at the boundary $u=0$. Hence

$$(-\nabla^2 u,\ u)=\int_{\Omega}(\nabla u)^2 d\Omega\geq 0. \tag{4}$$

It remains to prove that $u\equiv 0$ if $(-\nabla^2 u,\ u)=0$. This is obvious: if $(-\nabla^2 u,\ u)=0$ then, as follows from (4), $\nabla u\equiv 0$, whence it follows that u is constant. Being equal to zero at S this constant equals zero.

From what has been said it follows that the Dirichlet problem that we have formulated is equivalent to the problem of minimizing the functional

$$F(u)=(-\nabla^2 u,\ u)-2(u,\ f)$$

or, making use of formula (3),

$$F(u)=\int_{\Omega}\{(\nabla u)^2-2uf\}d\Omega. \tag{5}$$

As is known, the problem of the deflexion of a membrane fixed at its edge under the influence of a normal load reduces to equation (1) with boundary conditions (2), the function $f(P)$ being proportional to the loading. Functional (5) is

proportional to the potential energy of the curved membrane (see E. Trefftz [2] No. 26). We are thus concerned here with a particular case of the principle of minimum potential energy.

A similar result is obtained if in place of (2) we substitute the boundary condition of the mixed problem:

$$\left[\frac{\partial u}{\partial n}+\sigma(P)u\right]_S=0, \tag{6}$$

where $\sigma(P)$ is a non-negative function, not identically zero.

This time we must introduce, as the field of definition of the Laplacian operator, the linear set M_σ of functions which satisfy boundary conditions (6) and at the same time the same conditions of continuity and differentiability as the functions of the linear set M. It is easily seen that for M_σ the operator $-\nabla^2 u$ is also positive definite. In fact, by formula (3) and boundary condition (6),

$$(-\nabla^2 u,\, u)=-\int_S u\frac{\partial u}{\partial n}\,\mathrm{d}S+\int_\Omega(\nabla u)^2\mathrm{d}\Omega$$

$$=\int_S \sigma u^2\mathrm{d}S+\int_\Omega(\nabla u)^2\mathrm{d}\Omega\geq 0;$$

in which case, if $(-\nabla^2 u,\, u)=0$, then necessarily

$$\int_S \sigma u^2\mathrm{d}S=0,\quad \int_\Omega(\nabla u)^2\mathrm{d}\Omega=0.$$

From the second equation it follows that $u=C=\text{const}$. Substituting this in the first equation we obtain

$$C^2\int_S \sigma\mathrm{d}S=0.$$

The function $\sigma(P)$ is non-negative, moreover $\sigma(P)$ is not identically equal to zero. Hence it follows that $\int_S \sigma\mathrm{d}S>0$, but then necessarily $C=0$ and $u(P)\equiv 0$. Thus for the linear

set M_σ the operator $-\nabla^2$ is positive definite. The problem of integrating the Poisson equation (1) under boundary condition (6) can be replaced by the problem of minimizing the functional

$$F(u)=(-\nabla^2 u,\ u)-2(u,\ f)=\int_\Omega\{(\nabla u)^2-2uf\}\mathrm{d}\Omega+\int_S\sigma u^2\mathrm{d}S \quad (7)$$

for the linear set M_σ. Boundary condition (6) is natural.

We consider specially the Neumann problem for equation (1). We require to find the solution of equation (1) which is continuous and continuously differentiable in $\overline{\Omega}$ and which satisfies the boundary condition

$$\left.\frac{\partial u}{\partial n}\right|_S=0. \quad (8)$$

We can distinguish, analogously to the foregoing, a linear set M_0 consisting of functions which satisfy boundary condition (8) and the same conditions of continuity and differentiability as the functions of the linear set M.

However, the operator $-\nabla^2 u$ will not be positive definite for this lincar set. In fact, by formulae (3) and (8),

$$(-\nabla^2 u,\ u)=\int_\Omega(\nabla u)^2\mathrm{d}\Omega,$$

which represents a non-negative quantity. But from the equality $(-\nabla^2 u,\ u)=0$ it does not follow that $u\equiv 0$. Indeed the function $u\equiv 1$ clearly belongs to the linear set M_0 and $(-\nabla^2 u,\ u)=0$.

To obviate this difficulty, we proceed in the following manner. First of all we note that the Neumann problem is insoluble for an arbitrary function $f(P)$ and that it is easy to find the conditions which this function must necessarily satisfy. We integrate (1) over Ω:

$$-\int_\Omega\nabla^2 u\mathrm{d}\Omega=\int_\Omega f\mathrm{d}\Omega;$$

putting $v \equiv 1$ in formula (8) § 2, we obtain

$$\int_{\Omega} \nabla^2 u \, d\Omega = \int_{S} \frac{\partial u}{\partial n} \, dS,$$

which equals zero in virtue of (8).

Thus for the solubility of the Neumann problem it is necessary that

$$\int_{\Omega} f(P) d\Omega = 0.$$

Further, if the Neumann problem is soluble then it has an infinite set of solutions all differing from one another by a constant. We choose the solution $u(P)$ which satisfies the same condition as the given function $f(P)$

$$\int_{\Omega} u(P) d\Omega = 0. \tag{9}$$

Obviously such a solution of the Neumann problem is unique.

Considering the Neumann problem for equation (1), we adopt as the field of definition of the Laplacian operator the linear set of functions which: 1) are continuous together with their first and second derivatives in $\overline{\Omega}$, 2) satisfy boundary condition (8), 3) satisfy equation (9). We denote this set by $\tilde{M}_0$. It was stated above that each time we shall seek a solution of the Neumann problem satisfying condition (9). This can be formulated by saying that we shall seek the solution of the Neumann problem belonging to the linear set $\tilde{M}_0$. We shall prove that for the linear set $\tilde{M}_0$ the operator $-\nabla^2$ is positive.

In fact, we find as before

$$(-\nabla^2 u, u) = \int_{\Omega} (\nabla u)^2 d\Omega \geq 0.$$

If $(-\nabla^2 u, u) = 0$ then $u = \text{const}$, but a constant satisfying equality (9) is necessarily equal to zero.

This Neumann problem can be replaced by the following variational problem: to find in the linear set $\tilde{M}_0$ the function minimizing the functional

$$F(u) = (-\nabla^2 u, u) - 2(f, u)$$

or in virtue of boundary condition (8),

$$F(u) = \int_\Omega \{(\nabla u)^2 - 2fu\} d\Omega. \tag{10}$$

Boundary condition (8) is natural and so in seeking the minimum of functional (10), there is no need to satisfy this boundary condition beforehand.

It is often required to solve the Laplace equation with non-homogeneous boundary conditions accompanying it.

We have already seen (§ 18) that the integration of the Laplace equation in the domain Ω under the boundary condition (of the Dirichlet problem)

$$u|_S = g(P) \tag{11}$$

reduces to the finding of the minimum of the functional

$$\int_\Omega (\nabla u)^2 d\Omega \tag{12}$$

with respect to the set of functions satisfying condition (11). If the boundary condition has the form (of the Neumann problem)

$$\left.\frac{\partial u}{\partial n}\right|_S = h(P), \tag{13}$$

then the problem reduces to the finding of the minimum of the functional

$$\int_\Omega (\nabla u)^2 d\Omega - 2\int_S uh dS \tag{14}$$

with respect to the set of functions not subject to any boundary conditions. We now add to this that in the case of boundary conditions of the mixed type

$$\left[\frac{\partial u}{\partial n} + \sigma(P)u\right]_S = h_1(P) \tag{15}$$

the corresponding functional has the form

$$\int_{\Omega}(\nabla u)^2 d\Omega + \int_{S}(\sigma u^2 - 2uh_1)dS. \tag{16}$$

We may easily convince ourselves of this by reasoning similar to that of § 18.

Above we established the positive definite character of the operator $-\nabla^2$ for each of the sets M, M_σ and $\tilde{M}_0$. However, in order to be able to use the methods expounded in Ch. III with assurance it is important to establish that the operator mentioned is positive-bounded-below for each of the sets enumerated. We shall now go into this.

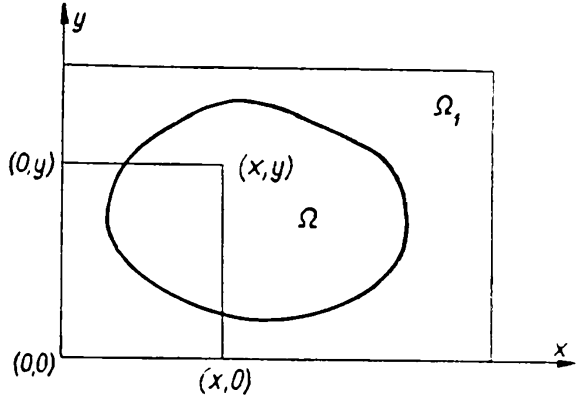

FIG. 8

We start with the Dirichlet problem. In this case

$$(-\nabla^2 u, u) = \int_{\Omega}(\nabla u)^2 d\Omega = \int_{\Omega}\sum_{k=1}^{m}\left(\frac{\partial u}{\partial x_k}\right)^2 d\Omega \tag{17}$$

where m is the number of dimensions of the domain Ω. For simplicity in calculation we shall assume that Ω is part of the xy-plane; the extension to any number of dimensions presents no difficulty.

For $m=2$ formula (17) assumes the simpler form

$$(-\nabla^2 u,\ u)=\int\limits_{\Omega}\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]\mathrm{d}\Omega.$$

We further recall that the region Ω is assumed to be finite.

We enclose Ω within some rectangle Ω_1, and take the coordinate axes along two of its sides. We denote the lengths of the sides of the rectangle by a and b (Fig. 8). Let $u(x, y)$ be a sufficiently smooth function, equal to zero at S.

(By a "sufficiently smooth" function we understand a function with a sufficient number of derivatives continuous in the closed region. In the given case the continuity of the first derivatives of the function $u(x, y)$ is sufficient.)

We extend $u(x, y)$ over the entire rectangle by taking it equal to zero outside Ω; with such an extrapolation it will be continuous in Ω_1. We take an arbitrary point (x_1, y_1) in Ω_1. We have, obviously,

$$\int\limits_0^{x_1}\frac{\partial u(x, y_1)}{\partial x}\mathrm{d}x=u(x_1, y_1)-u(0, y_1).$$

But the point $(0, y_1)$ lies outside Ω and thus $u(0, y_1)=0$. Thus

$$u(x_1, y_1)=\int\limits_0^{x_1}\frac{\partial u(x, y_1)}{\partial x}\mathrm{d}x.$$

Applying Cauchy's inequality, we have

$$u^2(x_1, y_1)\leq x_1\int\limits_0^{x_1}\left[\frac{\partial u(x, y_1)}{\partial x}\right]^2\mathrm{d}x\leq a\int\limits_0^{a}\left[\frac{\partial u(x, y_1)}{\partial x}\right]^2\mathrm{d}x.$$

We now integrate this inequality over the range $0\leq x_1\leq a$, $0\leq y_1\leq b$:

$$\int\limits_{\Omega_1}u^2(x_1, y_1)\mathrm{d}x_1\mathrm{d}y_1\leq a^2\int\limits_{\Omega_1}\left(\frac{\partial u}{\partial x}\right)^2\mathrm{d}x\mathrm{d}y\leq a^2\int\limits_{\Omega_1}\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]\mathrm{d}x\mathrm{d}y.$$

It is sufficient to take the integration over Ω, since $u \equiv 0$ outside Ω writing $a^2 = \frac{1}{\varkappa}$ we arrive at the important inequality

$$\int_{\Omega}\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]\mathrm{d}\Omega \geq \varkappa \int_{\Omega} u^2 \mathrm{d}\Omega, \tag{18}$$

which is known as *Friedrichs's inequality* [1]. It is valid in all cases, as we have established where the function $u(x, y)$ is continuously differentiable and equal to zero on S. In the general case of an m-dimensional space the Friedrichs inequality has the form

$$\int_{\Omega}\sum_{k=1}^{m}\left(\frac{\partial u}{\partial x_k}\right)^2 \mathrm{d}\Omega \geq \varkappa \int_{\Omega} u^2 \mathrm{d}\Omega, \qquad u|_S = 0; \tag{18$_1$}$$

For $m=1$ it transforms into inequality (8) § 6.

From inequality (18) it follows that

$$(-\nabla^2 u, u) \geq \gamma^2 \|u\|^2, \qquad \gamma = \sqrt{\varkappa} > 0, \tag{19}$$

whence it follows that the operator $-\nabla^2$ is positive-bounded -below for the linear set M of sufficiently smooth functions, equal to zero on S.

We now consider the operator $-\nabla^2 u$ for the linear set of functions satisfying boundary condition (6). The proof (that the operator is positive-bounded-below) in this case rests upon an inequality also established by Friedrichs:

$$\int_{\Omega} u^2 \mathrm{d}\Omega \leq C\left\{\int_{\Omega}\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]\mathrm{d}\Omega + \int_{S} u^2 \mathrm{d}S\right\}, \tag{20}$$

where C is a positive constant. In order to obtain this inequality we proceed in the following fashion. We again enclose Ω within a rectangle Ω_1 and put $u = fv$, where f is a function which we define later. From the easily verified identity

$$\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2 = f^2\left[\left(\frac{\partial v}{\partial x}\right)^2+\left(\frac{\partial v}{\partial y}\right)^2\right] - v^2 f \nabla^2 f + \frac{\partial}{\partial x}\left(v^2 f \frac{\partial f}{\partial x}\right) + \frac{\partial}{\partial y}\left(v^2 f \frac{\partial f}{\partial y}\right)$$

we find by dropping the first term on the right,

$$\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2 \geq -v^2 f\nabla^2 f+\frac{\partial}{\partial x}\left(v^2 f\frac{\partial f}{\partial x}\right)+\frac{\partial}{\partial y}\left(v^2 f\frac{\partial f}{\partial y}\right).$$

We integrate this over Ω and transform the integrals of the first and third terms on the right by Gauss's theorem (§ 2):

$$\int_\Omega\left\{\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right\}\mathrm{d}\Omega \geq -\int_\Omega v^2 f\nabla^2 f\mathrm{d}\Omega+\int_S v^2 f\frac{\partial f}{\partial n}\,\mathrm{d}S.$$

Hence

$$-\int_\Omega v^2 f\nabla^2 f\mathrm{d}\Omega \leq \int_\Omega\left\{\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right\}\mathrm{d}\Omega-\int_S v^2 f\frac{\partial f}{\partial n}\,\mathrm{d}S$$

and, *a fortiori*

$$-\int_\Omega v^2 f\,\nabla^2 f\mathrm{d}\Omega \leq \int_\Omega\left\{\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right\}\mathrm{d}\Omega+\left|\int_S v^2 f\frac{\partial f}{\partial n}\mathrm{d}S\right|.$$

We take $f=\sin\frac{\pi x}{a}\sin\frac{\pi y}{b}$. Then $\nabla^2 f=-\pi^2\left(\frac{1}{a^2}+\frac{1}{b^2}\right)f$ and f does not vanish anywhere in Ω. In the last inequality, the integral on the left becomes

$$\pi^2\left(\frac{1}{a^2}+\frac{1}{b^2}\right)\int_\Omega u^2\mathrm{d}\Omega. \tag{21}$$

The second integral on the right is evaluated thus:

$$\left|\int_S v^2 f\frac{\partial f}{\partial n}\,\mathrm{d}S\right|=\left|\int_S u^2\frac{1}{f}\frac{\partial f}{\partial n}\,\mathrm{d}S\right| \leq \int_S u^2\frac{1}{f}\left|\frac{\partial f}{\partial n}\right|\mathrm{d}S.$$

On the contour S the quantity $\frac{1}{f}\left|\frac{\partial f}{\partial n}\right|$ is obviously bounded. Let $\frac{1}{f}\left|\frac{\partial f}{\partial n}\right| \leq C'$ (= const) on S. Then

$$\left|\int_S u^2 f\frac{\partial f}{\partial n}\,\mathrm{d}S\right| \leq C'\int_S u^2\mathrm{d}S.$$

Inequality (20) follows from (21) if by C is denoted the least of the numbers $\pi^{-2}\left(\frac{1}{a^2}+\frac{1}{b^2}\right)^{-1}$ and $C'\pi^{-2}\left(\frac{1}{a^2}+\frac{1}{b^2}\right)^{-2}$.

We now estimate the scalar product $(-\nabla^2 u,\ u)$. We have

$$(-\nabla^2 u,\ u) = -\int_{\Omega} u\nabla^2 u \, d\Omega$$

$$= \int_{\Omega} \left\{ \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 \right\} d\Omega - \int_{\Omega} u \frac{\partial u}{\partial n} \, dS,$$

or if we use equation (6)

$$(-\nabla^2 u,\ u) = \int_{\Omega} \left\{ \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 \right\} d\Omega + \int_{S} \sigma u^2 dS$$

$$\geq \int_{\Omega} \left\{ \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 \right\} d\Omega + \sigma_0 \int_{S} u^2 dS,$$

where, as has already been stated, σ_0 is the lower bound of $\sigma(P)$. We put $\sigma_1 = \min(\sigma_0, 1)$; then obviously

$$(-\nabla^2 u,\ u) \geq \sigma_1 \left\{ \int_{\Omega} \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 \right] d\Omega + \int_{S} u^2 dS \right\}. \quad (22)$$

Comparing this with (20) we finally obtain

$$(-\nabla^2 u,\ u) \geq \gamma^2 \|u\|^2, \qquad \gamma = \sqrt{\left(\frac{\sigma_1}{C}\right)}. \quad (23)$$

From what has been proved by virtue of the theorems of §§ 11, 15 and 19, it follows that the problem of finding a minimum to which the integration of the Poisson equation under the boundary condition (2) or (6) reduces, is soluble and can be solved approximately by the Ritz method or by one of the methods of Chap. III.

We note that, for the Poisson equation, energy convergence leads in the Dirichlet problem [condition (2)] to convergence in the mean over the region Ω both of the approximate solutions themselves and of their first derivatives. In the

mixed problem [condition (6)] there is in addition convergence in the mean of the approximate solutions over the boundary S. This is easily seen from formulae (17) and (22).

(It follows from S. L. Sobolev's "embedded theorems" [2] that convergence in the mean of functions over the boundary is a consequence of convergence in the mean of the first derivatives within the domain.)

We now turn to the Neumann problem. The well-known *Poincaré inequality*

$$\int_{\Omega} u^2 d\Omega \leq A \int_{\Omega} \left\{ \left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right\} d\Omega + B \left(\int_{\Omega} u d\Omega \right)^2, \quad (24)$$

where $A > 0$ and $B > 0$ are constants, plays an essential role in the investigation of this problem. We present the proof of this inequality for the case when the domain Ω is a rectangle; the proof for domains of an arbitrary form can be found, for example, in the book by Courant and Friedrichs [2] Ch. VII, and also in S. L. Sobolev's book [2].

Let Ω be the rectangle $0 \leq x \leq a$, $0 \leq y \leq b$. We take two points (x_1, y_1) and (x_2, y_2) in Ω. We have, obviously,

$$u(x_2, y_2) - u(x_1, y_1) = \int_{x_1}^{x_2} \frac{\partial u(x, y_1)}{\partial x} dx + \int_{y_1}^{y_2} \frac{\partial u(x_2, y)}{\partial y} dy.$$

Squaring this and using the Cauchy inequality, we obtain

$$\begin{aligned} & u^2(x_1, y_1) + u^2(x_2, y_2) - 2u(x_1, y_1)u(x_2, y_2) \\ & \leq 2 \left\{ \left(\int_{x_1}^{x_2} \frac{\partial u(x, y_1)}{\partial x} dx \right)^2 + \left(\int_{y_1}^{y_2} \frac{\partial u(x_2, y)}{\partial y} dy \right)^2 \right\} \\ & \leq 2 \left\{ |x_2 - x_1| \int_0^a \left(\frac{\partial u(x, y_1)}{\partial x} \right)^2 dx + |y_2 - y_1| \int_0^b \left(\frac{\partial u(x_2, y)}{\partial y} \right)^2 dy \right\} \\ & \leq 2 \left\{ a \int_0^a \left(\frac{\partial u(x, y_1)}{\partial x} \right)^2 dx + b \int_0^b \left(\frac{\partial u(x_2, y)}{\partial y} \right)^2 dy \right\}. \end{aligned}$$

We integrate the inequality obtained with respect to each of x_1, y_1, x_2, y_2 over $0 \leq x_1, x_2 \leq a$, $0 \leq y_1, y_2 \leq b$.
We then obtain

$$2ab \int_{\Omega} u^2 \mathrm{d}\Omega - 2 \left(\int_{\Omega} u \mathrm{d}\Omega \right)^2$$

$$\leq 2ab \left\{ a^2 \int_{\Omega} \left(\frac{\partial u}{\partial x} \right)^2 \mathrm{d}\Omega + b^2 \int_{\Omega} \left(\frac{\partial u}{\partial y} \right)^2 \mathrm{d}\Omega \right\}.$$

Dividing by $2ab$ and putting

$$A = \max (a^2, b^2), \quad B = \frac{1}{ab},$$

we arrive at the Poincaré inequality for a rectangle. The extension for the case of a greater number of dimensions encounters no difficulties.

As soon as the Poincaré inequality is established the proof that the operator $-\nabla^2$ is positive-bounded-below for the set $\tilde{M}_0$ is made easy. We have already seen above that for functions of this set

$$(-\nabla^2 u, u) = \int_{\Omega} (\nabla u)^2 \mathrm{d}\Omega.$$

Since the function $u(P)$ satisfies the further condition (9) the Poincaré inequality simplifies and assumes the form

$$\int_{\Omega} u^2 \mathrm{d}\Omega \leq A \int_{\Omega} \sum_{k=1}^{m} \left(\frac{\partial u}{\partial x_k} \right)^2 \mathrm{d}\Omega = A \int_{\Omega} (\nabla u)^2 \mathrm{d}\Omega.$$

Now it is clear that

$$(-\nabla^2 u, u) \geq \gamma^2 \|u\|^2, \quad \gamma = \frac{1}{\sqrt{A}},$$

which was to be proved. From what has been proved there follows both the solubility of the Neumann problem for

equation (1) provided the function $f(P)$ has a finite norm and satisfies equation (9), and also the possibility of finding an approximate solution by one of the methods of ch. III and, in particular, by the Ritz method.

§ 23. Problems of torsion of a bar and of bending of a beam by a lateral force

We assume that the reader is familiar with the theories of torsion and bending of rods, e.g. from the course by Leibenzon [2] or by S. P. Timoshenko [1]. The application of the energy method to the problems of bending and torsion is frequently met with in L. S. Leibenzon's [1] monograph where there are several of the formulae presented here and a large number of examples are considered.

We denote by Ω the normal cross-section of the rod and by S the boundary of the section. In the general case the problem of torsion reduces to the finding of a function $\varphi_0(x,y)$ which satisfies the Laplace equation

$$\nabla^2\varphi_0=\frac{\partial^2\varphi_0}{\partial x^2}+\frac{\partial^2\varphi_0}{\partial y^2}=0, \tag{1}$$

over the cross-section, and on the boundary S the boundary condition

$$\frac{\partial\varphi_0}{\partial n}=ly-mx. \tag{2}$$

Here n is the outward drawn normal from S, $l=\cos(n,z)$, $m=\cos(n,y)$

Thus the function $\varphi_0(x,y)$ is the solution of the Neumann problem for the Laplace equation under the non-homogeneous boundary condition (2).

Applying those results of the previous section which are relevant to this problem, we find that the torsion problem

can be reduced to the finding of a function which minimizes the functional

$$\int_{\Omega}\int(\nabla\varphi)^2 d\Omega - 2\int_{S}\varphi(ly-mx)dS$$
$$=\int_{\Omega}\int\left[\left(\frac{\partial\varphi}{\partial x}\right)^2+\left(\frac{\partial\varphi}{\partial y}\right)^2\right]dxdy-2\int_{S}\varphi(ly-mx)dS. \qquad (3)$$

The functions $\varphi(x,y)$ are not subjected beforehand to any boundary conditions, since condition (2) is natural for the functional (3).

We transform the line integral in expression (3) by Stokes's theorem.

$$\int_{S}[\varphi y\cos(n,x)-\varphi x\cos(n,y)]dS=\int_{\Omega}\int\left[\frac{\partial(\varphi y)}{\partial x}-\frac{\partial(\varphi x)}{\partial y}\right]dxdy$$
$$=\int_{\Omega}\int\left(y\frac{\partial\varphi}{\partial x}-x\frac{\partial\varphi}{\partial y}\right)dxdy.$$

Substituting this in (3) we readily bring our functional to the form

$$\int_{\Omega}\int\left[\left(\frac{\partial\varphi}{\partial x}-y\right)^2+\left(\frac{\partial\varphi}{\partial y}+x\right)^2\right]dxdy-\int_{\Omega}\int(x^2+y^2)dxdy.$$

The second integral, equal to the moment of inertia of the section about the coordinate origin is a constant quantity, therefore we need only seek the minimum of a simpler functional which we shall denote by $D(\varphi)$:

$$D(\varphi)=\int_{\Omega}\int\left[\left(\frac{\partial\varphi}{\partial x}-y\right)^2+\left(\frac{\partial\varphi}{\partial y}+x\right)^2\right]dxdy. \qquad (4)$$

The functional $D(\varphi)$ has a simple mechanical meaning. It is well known that

$$D(\varphi_0)=\frac{C}{G}, \qquad (5)$$

where C is the torsional rigidity of the rod, G is the shear modulus of the rod material and φ_0, as was stated above, is the function solving the torsion problem.

The proof of formula (5) is simple. We have (see, e.g. L. S. Leibenzon [2], p. 242)

$$\frac{C}{G} = \int_\Omega \int \left(x^2 + y^2 + x\frac{\partial \varphi_0}{\partial y} - y\frac{\partial \varphi_0}{\partial x} \right) dx dy.$$

On the other hand, expanding the brackets, we easily find

$$D(\varphi_0) = \frac{C}{G} + \int_\Omega \int \left[\left(\frac{\partial \varphi_0}{\partial x} \right)^2 + \left(\frac{\partial \varphi_0}{\partial y} \right)^2 - y\frac{\partial \varphi_0}{\partial x} + x\frac{\partial \varphi_0}{\partial x} \right] dx dy.$$

By Green's formula

$$\int_\Omega \int \left[\left(\frac{\partial \varphi_0}{\partial x} \right)^2 + \left(\frac{\partial \varphi_0}{\partial y} \right)^2 \right] dx dy = - \int_\Omega \int \varphi_0 \nabla^2 \varphi_0 \, dx dy + \int_S \varphi_0 \frac{\partial \varphi_0}{\partial n} dS,$$

or by virtue of equations (1) and (2),

$$\int_\Omega \int \left[\left(\frac{\partial \varphi_0}{\partial x} \right)^2 + \left(\frac{\partial \varphi_0}{\partial y} \right)^2 \right] dx dy = \int_S \varphi_0 (ly - mx) dS,$$

which, as we have already seen, equals

$$\int_\Omega \int \left(y\frac{\partial \varphi_0}{\partial x} - x\frac{\partial \varphi_0}{\partial y} \right) dx dy.$$

Substituting this in $D(\varphi_0)$, we obtain formula (5).

The function φ_0 minimizes the functional $D(\varphi)$ and so, no matter what the function $\varphi(x,y)$ is, there follows the well-known inequality

$$C \leq G \int_\Omega \int \left[\left(\frac{\partial \varphi}{\partial x} - y \right)^2 + \left(\frac{\partial \varphi}{\partial y} + x \right)^2 \right] dx dy. \tag{6}$$

This inequality enables us to estimate a bound to the torsional rigidity. If as φ we take a function approximately solving the problem of the minimum of the function $D(\varphi)$

(e.g. by constructing this function by the Ritz method), then the estimate given by formula (6) can be as exact as we please. We obtain a simple, albeit rough, estimate for the torsional rigidity by putting $\varphi \equiv 0$. This gives the inequality

$$C \leq GI, \tag{7}$$

where I is the moment of inertia of the rod cross-section about the coordinate origin.

It is possible to obtain a rather more precise estimate of the rigidity C. To this end we put $\varphi = \alpha xy$ in (6), where we choose the constant α such that the right-hand side of inequality (6) assumes its least value. Extremely simple manipulation leads to the estimate

$$C \leq \frac{4GI_xI_y}{I}, \tag{7$_1$}$$

which is more precise than estimate (7). By I_x and I_y are denoted the geometrical moments of inertia of the cross-section of the rod about the axes Ox and Oy respectively.

The problem of torsion can be reduced to another variational problem which allows us to obtain a lower estimate for the torsional rigidity.

We consider, first of all, the simple case where the cross-section Ω is simply connected. Then, as is known, the torsion problem can be reduced to the integration of the Poisson equation:

$$\nabla^2 \psi = -2G\theta$$

under the boundary condition

$$\psi|_S = 0;$$

here θ is a constant (the degree of torsion). If we put $\psi(x, y) = 2G\theta u(x, y)$, we obtain the differential equation

$$-\nabla^2 u = 1 \tag{8}$$

and the boundary condition

$$u|_S = 0. \tag{9}$$

From the results of the previous section it follows that the problem (8) with (9) is equivalent to the problem of finding the minimum of the functional

$$F(u)=\int_{\Omega}\int\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2-2u\right]\mathrm{d}x\mathrm{d}y \tag{10}$$

with respect to the set of functions satisfying condition (9). Let $u_0(x, y)$ be the exact solution of our problem and $u_n(x, y)$ its Ritz approximation. By formula (15) § 14 $\|u_n\|^2\leq\|u_0\|^2$. In our problem $f(x, y)\equiv 1$, and formulae (13) and (14) § 14 give

$$\int_{\Omega}\int u_0(x, y)\mathrm{d}x\mathrm{d}y\geq\int_{\Omega}\int u_n(x, y)\mathrm{d}x\mathrm{d}y.$$

As is known, the torsional rigidity differs only by a factor $4G$ from the integral on the left. Hence there follows a lower estimate for C:

$$C\geq 4G\int_{\Omega}\int u_n(x, y)\mathrm{d}x\mathrm{d}y. \tag{11}$$

If we use formula (18) § 14 we obtain a, generally speaking, cruder estimate

$$C\geq\frac{4G\left(\int_{\Omega}\int u\mathrm{d}x\mathrm{d}y\right)^2}{\int_{\Omega}\int\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]\mathrm{d}x\mathrm{d}y}. \tag{12}$$

Here $u(x, y)$ is any function continuously differentiable in $\overline{\Omega}$ and vanishing on S.

Now let the cross-section Ω be multiply-connected and let the curve. S_0 be the external boundary and the curves S_1, S_2, ..., S_k, internal boundaries. We denote by γ_i, $i=1, 2$, ..., k the area enclosed within the curve S_i. The function $u(x, y)=\psi(x, y)/2G\theta$ satisfies equation (8) and the boundary conditions:

$$u=\begin{cases}0 & S_0,\\ \alpha_i & S_i,\quad i=1, 2, \ldots, k.\end{cases} \tag{13}$$

Here α_i are constants which are unknown beforehand; in virtue of the well-known Bredt theorem they are connected by the relations

$$\int_{S_i} \frac{\partial u}{\partial n}\, dS = -\gamma_i; \quad i=1,\ 2,\ \ldots,\ k; \tag{14}$$

n as usual, denotes the normal to S_i which is outward with respect to Ω and consequently inward with respect to the closed curve S_i.

Let $\tilde{u}(x,\ y)$ be some function satisfying conditions (13) and (14); we suppose further that in $\overline{\Omega}$ this function is continuous together with its first and second derivatives, We put $u-\tilde{u}=v$. Then the function v satisfies the equation

$$-\nabla^2 v = 1 + \nabla^2 \tilde{u} \tag{15}$$

and the boundary conditions

$$v = \begin{cases} 0 & S_0, \\ \beta_i & S_i, \quad i=1,\ 2,\ \ldots,\ k; \end{cases} \tag{16}$$

$$\int_{S_i} \frac{\partial v}{\partial n}\, dS = 0, \quad i=1,\ 2,\ \ldots,\ k. \tag{17}$$

Here the β_i are arbitrary constants. Functions satisfying conditions (16) and (17) constitute, as is easily seen, a linear set. We shall prove that for this linear set the operator $-\nabla^2$ is positive-definite. Indeed, if any function $v(x,\ y)$ satisfies the conditions mentioned, then

$$(-\nabla^2 v,\ v) = -\int_{\Omega}\int v \nabla^2 v\, dx\, dy$$

$$= \int_{\Omega}\int \left[\left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2\right] dx\, dy + \int_{S} v \frac{\partial v}{\partial n}\, dS.$$

The line integral vanishes, since

$$\int_{S} v \frac{\partial v}{\partial n}\, dS = \sum_{i=1}^{k} \beta_i \int_{S_i} \frac{\partial v}{\partial n}\, dS = 0$$

and

$$(-\nabla^2 v, v) = \int_\Omega \int \left[\left(\frac{\partial v}{\partial x} \right)^2 + \left(\frac{\partial v}{\partial x} \right)^2 \right] dx dy \geq 0.$$

The symmetry of our operatór may be proved analogously.

By the minimum functional theorem (§ 11) our task reduces to the finding of the function $v(x, y)$ which minimizes the functional

$$\int_\Omega \int \left[\left(\frac{\partial v}{\partial x} \right)^2 + \left(\frac{\partial v}{\partial y} \right)^2 - 2(1 + \nabla^2 \tilde{u}) v \right] dx dy \qquad (18)$$

with respect to the set of functions satisfying conditions (16) and (17).

We now turn to the function $u = v + \tilde{u}$. Dropping the constant terms in (18) we find that the integration of equation (8) under conditions (13) and (14) is equivalent to the finding of the function minimizing under these conditions, the functional

$$Q(u) = \int_\Omega \int \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 - 2(1 + \nabla^2 \tilde{u}) u \right] dx dy$$
$$- 2 \int_\Omega \int \left(\frac{\partial u}{\partial x} \frac{\partial \tilde{u}}{\partial x} + \frac{\partial u}{\partial y} \frac{\partial \tilde{u}}{\partial y} \right) dx dy.$$

We treat the second integral by Green's formula and use the boundary conditions:

$$\int_\Omega \int \left(\frac{\partial u}{\partial x} \frac{\partial \tilde{u}}{\partial x} + \frac{\partial u}{\partial y} \frac{\partial \tilde{u}}{\partial y} \right) dx dy = - \int_\Omega \int u \nabla^2 \tilde{u} \, dx dy + \int_S u \frac{\partial \tilde{u}}{\partial n} dS$$
$$= - \int_\Omega \int u \nabla^2 \tilde{u} \, dx dy + \sum_{i=1}^{k} \alpha_i \gamma_i.$$

Substituting this in $Q(u)$ we find the final expression for the functional $Q(u)$:

$$Q(u) = \int_\Omega \int \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 - 2u \right] dx dy - 2 \sum_{i=1}^{k} \alpha_i \gamma_i. \qquad (19)$$

It is easily proved by the criterion given in § 17 that condition (14) is natural for the functional $Q(u)$. Thus we look for its minimum using the set of functions satisfying conditions (13) alone. As before, we denote by $u_0(x, y)$ the function solving the torsion problem. It minimizes $Q(u)$. We now explain the connexion between $Q(u_0)$ and C, the torsional rigidity. We start from the well-known formula

$$C=4G\left\{\int_\Omega\int u_0\mathrm{d}x\mathrm{d}y+\sum_{i=1}^{k}\alpha_i^{(0)}\gamma_i\right\}; \tag{20}$$

by $\alpha_i^{(0)}$ are denoted the values of the constants α_i for the function u_0. Since $\nabla^2 u_0=-1$, the integral in (20) transforms in the following way:

$$\int_\Omega\int u_0\mathrm{d}x\mathrm{d}y=-\int_\Omega\int u_0\nabla^2 u_0\mathrm{d}x\mathrm{d}y=\int_\Omega\int\left[\left(\frac{\partial u_0}{\partial x}\right)^2+\left(\frac{\partial u_0}{\partial y}\right)^2\right]\mathrm{d}x\mathrm{d}y$$

$$-\int_S u_0\frac{\partial u_0}{\partial n}\mathrm{d}S=\int_\Omega\int\left[\left(\frac{\partial u_0}{\partial x}\right)^2+\left(\frac{\partial u_0}{\partial y}\right)^2\right]\mathrm{d}x\mathrm{d}y-\sum_{i=1}^{k}\alpha_i^{(0)}\gamma_i.$$

Hence

$$\int_\Omega\int\left[\left(\frac{\partial u_0}{\partial x}\right)^2+\left(\frac{\partial u_0}{\partial y}\right)^2\right]\mathrm{d}x\mathrm{d}y=\int_\Omega\int u_0\mathrm{d}x\mathrm{d}y+\sum_{i=1}^{k}\alpha_i^{(0)}\gamma\,.$$

Substituting u_0 for u in (19) and using the last equation we find

$$Q(u_0)=-\int_\Omega\int u_0\mathrm{d}x\mathrm{d}y-\sum_{i=1}^{k}\alpha_i^{(0)}\gamma_i.$$

Comparing this with (20) we obtain

$$C=-4GQ(u_0). \tag{21}$$

Let u be any function satisfying conditions (13). We select the constant factor a such that $Q(au)$ is a minimum. We have

$$Q(au)=a^2\int_\Omega\int\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]\mathrm{d}x\mathrm{d}y-$$

$$-2a\left[\int_{\Omega}\int u\,dx\,dy+\sum_{i=1}^{k}\alpha_i\gamma_i\right].$$

Differentiating this with respect to a and equating the derivative to zero, we find

$$a=\frac{\int_{\Omega}\int u\,dx\,dy+\sum_{i=1}^{k}\alpha_i\gamma_i}{\int_{\Omega}\int\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]dx\,dy}.$$

It is now easy to show that

$$Q(au)=-\frac{\left(\int_{\Omega}\int u\,dx\,dy+\sum_{i=1}^{k}\alpha_i\gamma_i\right)^2}{\int_{\Omega}\int\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]dx\,dy}. \tag{22}$$

Obviously $Q(u_0)\leq Q(au)$ since au satisfies conditions (13) (in which the constants α_i are replaced by $a\alpha_i$). There now follows a lower estimate for the torsional rigidity from formulae (21) and (22)

$$C\geq 4G\frac{\left(\int_{\Omega}\int u\,dx\,dy+\sum_{i=1}^{k}\alpha_i\gamma_i\right)^2}{\int_{\Omega}\int\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]dx\,dy}. \tag{23}$$

Estimate (23) was cited in the paper by Lin Hung-Sun [1].

We now briefly mention the problem of bending of a rod. It reduces as is known, to the finding of a function $\chi(x, y)$, which is harmonic in the region Ω and satisfies the boundary condition

$$\left.\frac{\partial\chi}{\partial n}\right|_s=-\left[\frac{1}{2}\sigma x^2+\left(1-\frac{\sigma}{2}\right)y^2\right]l-(2+\sigma)xym; \tag{24}$$

σ is Poisson's ratio. According to what was proved in the previous paragraph, this problem reduces to determining a function which minimizes functional

$$\int_{\Omega}\int \left[\left(\frac{\partial \chi}{\partial x}\right)^2+\left(\frac{\partial \chi}{\partial y}\right)^2\right]\mathrm{d}x\mathrm{d}y$$

$$+\int_{S}\chi\{[\sigma x^2+(2-\sigma)y^2]l+2(2+\sigma)xym\}\mathrm{d}S.$$

Since condition (24) is natural, no boundary conditions need be imposed upon the function χ.

§ 24. Equations with variable coefficients

An important role is played in various questions of mathematical physics by differential equations with variable coefficients, that is linear equations, the coefficients of which are given functions of the independent variables of the problem, particularly of the coordinates. Thus, the equation for the equilibrium of and also for the oscillations of an elastic envelope is a system of equations with variable coefficients. The deflexion of a plate of variable thickness is defined by certain fourth-order equations with variable coefficients. The problem of the oscillations of the surface of a liquid in a vessel reduces to a second-order equation with variable coefficients (see e.g. L. Sretensky [1]). Linear equations with variable coefficients are of special importance for the investigations of non-linear equations. On the one hand the various methods of linearization often lead to equations with variable coefficients, for instance, the derivation of an important equation such as that of Chaplygin in gas dynamics [1]. On the other hand, the application of several methods of successive approximation to non-linear equations often results in each approximation being obtained as the solution of a certain linear equation with variable coefficients. This often happens

during the use of the so called "Newton's method" elaborated by L.V. Kantorovich [1].

The transition from constant to variable coefficients leads to serious difficulty in the use of such methods as the method of integral equations or the method of Green's function, and especially in finding elementary particular solutions. But for the majority of variational methods (including the energy method) the transition to variable coefficients is not associated with any significant increase in difficulty and the results of § 22 apply *mutatis mutandis* to equations of elliptic type with variable coefficients as defined later. The present section is devoted to this question.

We shall consider the equation

$$Lu = -\sum_{j,k=1}^{m} \frac{\partial}{\partial x_j}\left(A_{jk}(P)\frac{\partial u}{\partial x_k}\right) + C(P)u = f(P); \quad A_{jk} = A_{kj} \tag{1}$$

under one of the three types of boundary conditions

$$u|_S = 0; \tag{2}$$

$$[N(u) + \sigma(P)u]_S = 0, \tag{3}$$

$$N(u)|_S = 0, \tag{4}$$

where $$N(u) = \sum_{j,k=1}^{m} A_{jk}(P)\frac{\partial u}{\partial x_k}\cos(n, x_j);$$

where, as usual, S is the boundary of some finite domain Ω, and n is the outward drawn normal to S. By analogy with the cases of the Poisson or Laplace equations, the boundary value problems for equation (1) under conditions (2), (3) or (4) are known respectively as the Dirichlet, mixed, or Neumann problem. The choice of the special form of equation (1) (the general second-order equation does not reduce to this form) is explained by the fact that under any of boundary conditions (2), (3), or (4) the operator L is symmetric. In fact, by Green's formula (formula (7) § 2)

$$\int_\Omega (vLu - uLv)\,d\Omega = -\int_S [vN(u) - uN(v)]\,dS. \tag{5}$$

If both functions satisfy condition (2) or condition (4), the function under the sign of the surface integral obviously vanishes. If both functions satisfy condition (3) then on S

$$N(u)+\sigma u=0, \qquad N(v)+\sigma v=0.$$

Multiplying the first equation by v, the second by u and subtracting, we find

$$[vN(u)-uN(v)]_S=0,$$

so that in this case the function under the sign of the integral over S vanishes. Thus in all cases the right-hand side and, consequently, also the left-hand side of equation (15) vanishes:

$$\int_{\Omega}(vLu-uLv)\mathrm{d}\Omega=(Lu,\ v)-(u,\ Lv)=0,$$

and this means that the operator L is symmetric for the corresponding sets of functions.

The operator L and equation (1) are known as *elliptic in the closed domain* $\overline{\Omega}$ if the coefficients $A_{jk}(P)$ possess the following property: there exists a positive constant μ_0 such that, for no matter what real numbers $t_1, t_2, \ldots, t_m$ and for all points $P\in\overline{\Omega}$, the inequality

$$\sum_{j,\,k=1}^{m} A_{jk}(P)t_jt_k\geq\mu_0\sum_{i=1}^{m}t_k^2 \tag{6}$$

holds, no restrictions being imposed upon the coefficient C.

This definition also holds true for the general linear second-order differential operator

$$-\sum_{j,\,k=1}^{m}A_{jk}\frac{\partial^2u}{\partial x_i\partial x_j}+\sum_{i=1}^{m}B_i\frac{\partial u}{\partial x_i}+Cu$$

and for the corresponding differential equation. The designation of elliptic in the closed region is naturally retained for the operator L if in the latter the sign of its coefficients is changed so that, instead of (6), we shall have the inequality

$$\sum_{j,\,k=1}^{m}A_{jk}(P)t_jt_k\leq-\mu_0\sum_{k=1}^{m}t_k^2.$$

We consider as an example the operator

$$\frac{\partial}{\partial x}\left[(1+y^2)\frac{\partial u}{\partial x}\right]-\frac{\partial}{\partial x}\left(xy\frac{\partial u}{\partial y}\right)-\frac{\partial}{\partial y}\left(xy\frac{\partial u}{\partial x}\right)+\frac{\partial}{\partial y}\left[(1+x^2)\frac{\partial u}{\partial y}\right].$$

Here $m=2$. If we put $x=x_1$, $y=x_2$, then

$$A_{11}=1+x_2^2, \quad A_{12}=A_{21}=-x_1x_2, \quad A_{22}=1+x_1^2.$$

The left-hand side of formula (6) in the given case has the form

$$(1+x_2^2)t_1^2-2x_1x_2t_1t_2+(1+x_1^2)t_2^2=t_1^2+t_2^2+(x_2t_1-x_1t_2)^2.$$

By dropping the last term on the right, we obtain

$$(1+x_2^2)t_1^2-2x_1x_2t_1t_2+(1+x_1^2)t_2^2\geq t_1^2+t_2^2.$$

Inequality (6) is satisfied when $\mu_0=1$ independently of the values of x_1, and x_2. Hence it follows that our operator is elliptic for all values of the variables x and y.

We consider further the "Tricomi operator"

$$y\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=x_2\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}.$$

Here $A_{11}=x_2$, $A_{12}=A_{21}=0$, $A_{22}=1$ and the left-hand side of formula (6) has the form $x_2t_1^2+t_2^2$. If $x_2\geq\delta>0$ then $x_2t_1^2+t_2^2 \geq \delta t_1^2+t_2^2\geq\delta'(t_1^2+t_2^2)$, where δ' denotes the least of the numbers 1 and δ; inequality (6) is satisfied when the constant $\mu_0=\delta$ and the operator $y\dfrac{\partial^2 u}{\partial x^2}+\dfrac{\partial^2 u}{\partial y^2}$ is elliptic in any closed domain which, together with its boundary, lies entirely within the upper half-plane. If the region lies entirely or partially within the lower half-plane, the Tricomi operator will not be elliptic in this domain being hyperbolic there. If the domain lies in the upper half-plane but its boundary has a common part (even just one common point) with the x-axis, then in the closed domain the Tricomi operator belongs to the class of so called *degenerate* elliptic operators which we shall discuss in the next section.

We shall assume that equation (1) is elliptic in $\overline{\Omega}$ so that the inequality (6) is satisfied for some positive constant μ_0. We shall explain the condition which ensures that the operator (1) is positive-bounded-below for each of the sets of functions satisfying the boundary condition (2), (3) or (4); as we know, this guarantees both the existence of a solution of the corresponding boundary value problem and the possibility of an approximate evaluation of this solution with the aid of the methods of Chapter III. In all cases it will be assumed that $C(P) \geq 0$. We shall not consider here the case when $C(P)$ can assume negative values.

By Green's formula [formula (5) § 2]

$$(Lu, u) = \int_{\Omega} uLu\,d\Omega = \int_{\Omega} \left\{ \sum_{j,k=1}^{m} A_{jk} \frac{\partial u}{\partial x_j} \frac{\partial u}{\partial x_k} + Cu^2 \right\} d\Omega - \int_{S} uN(u)\,dS.$$

If boundary condition (2) or (4) is given, then the surface integral vanishes and

$$(Lu, u) = \int_{\Omega} \left\{ \sum_{j,k=1}^{m} A_{jk} \frac{\partial u}{\partial x_j} \frac{\partial u}{\partial x_k} + Cu^2 \right\} d\Omega. \tag{7}$$

If condition (3) applies, then on S we shall have $N(u) = -\sigma u$ and

$$(Lu, u) = \int_{\Omega} \left\{ \sum_{j,k=1}^{m} A_{jk} \frac{\partial u}{\partial x_j} \frac{\partial u}{\partial x_k} + Cu^2 \right\} d\Omega + \int_{\Omega} \sigma u^2\,dS. \tag{8}$$

We shall assume that the function $\sigma(P)$ is bounded below by some positive number σ_0, so that $\sigma(P) \geq \sigma_0 > 0$.

If the coefficient $C(P)$ is bounded below by some positive number then the operator L is positive-bounded-below for any of the conditions (2) – (4).

In fact, by inequality (6)

$$\int_{\Omega} \sum_{j,k=1}^{m} A_{jk} \frac{\partial u}{\partial x_j} \frac{\partial u}{\partial x_k}\,d\Omega \geq \mu_0 \int_{\Omega} \sum_{k=1}^{m} \left(\frac{\partial u}{\partial x_k} \right)^2 d\Omega, \tag{9}$$

and the integral on the left in formula (9) is non-negative. In formulae (7) and (8) we drop this integral, and in formula (8) we further omit the non-negative surface integral. In both cases, we arrive at the inequality

$$(Lu, u) \geq \int_{\Omega} C(P)u^2 d\Omega.$$

By assumption, $C(P) \geq C_0$ where C_0 is some positive constant, but then from the last inequality it immediately follows that

$$(Lu, u) \geq \gamma^2 \int_{\Omega} u^2 d\Omega = \gamma^2 \|u\|^2; \qquad \gamma = \sqrt{(C_0)},$$

i.e. that L is a positive-bounded-below operator.

If the coefficient $C(P)$ can vanish somewhere in $\overline{\Omega}$ then the proof presented above is inapplicable, but the operator L remains positive-bounded-below for the boundary conditions (2) and (3). We consider the case of boundary condition (2). Under the integral sign in formula (7) we drop the non-negative term Cu^2 and estimate the remaining integral by inequality (9). We obtain as a result:

$$(Lu, u) \geq \mu_0 \int_{\Omega} \sum_{i=1}^{m} \left(\frac{\partial u}{\partial x_i} \right)^2 d\Omega.$$

The function $u(P)$ vanishes on the boundary of the region Ω and thus satisfies Friedrichs's inequality [formula (18_1) § 22]. Making use of this we obtain:

$$(Lu, u) \geq \gamma^2 \|u\|^2, \qquad \gamma = \sqrt{(\varkappa \mu_0)},$$

which was to be proved.

We now turn to condition (3). In formula (8) we drop the Cu^2 under the sign of the integral and under the sign of the surface integral we replace σ by σ_0. We estimate the remaining volume integral by inequality (9). All this gives

$$(Lu, u) \geq \mu_0 \int_{\Omega} \sum_{k=1}^{m} \left(\frac{\partial u}{\partial x_k} \right)^2 d\Omega + \sigma_0 \int_{S} u^2 dS.$$

Let α be the least of the numbers μ_0 and σ_0. Then, obviously,

$$(Lu, u) \geq \alpha \left\{ \int_\Omega \sum_{k=1}^{m} \left(\frac{\partial u}{\partial x_k} \right)^2 d\Omega + \int_S u^2 dS \right\}.$$

Now inequality (20) § 22, which can be simply generalized for the case of any m, gives

$$(Lu, u) \geq \gamma^2 \|u\|^2, \qquad \gamma = 1/(\alpha C).$$

Turning to boundary condition (4) we restrict ourselves to the simplest and at the same time least interesting case where $C(P) \equiv 0$. The problem reduces to the integration of the equation

$$L_0 u = -\sum_{j,k=1}^{m} \frac{\partial}{\partial x_j} \left(A_{jk} \frac{\partial u}{\partial x_k} \right) = f(P) \tag{10}$$

under boundary condition (4). It is easily seen that this problem does not always have a solution. In order to be convinced of this we integrate both sides of equation (10) over Ω:

$$-\int_\Omega \sum_{j,k=1}^{m} \frac{\partial}{\partial x_j} \left(A_{jk} \frac{\partial u}{\partial x_k} \right) d\Omega = \int_\Omega f d\Omega.$$

Applying Gauss's theorem to the left-hand side of the latter equation, we find

$$-\int_\Omega \sum_{j,k=1}^{m} \frac{\partial}{\partial x_j} \left(A_{jk} \frac{\partial u}{\partial x_k} \right) d\Omega = -\int_S \sum_{j,k=1}^{m} A_{jk} \frac{\partial u}{\partial x_k} \cos(n, x_j) dS = 0,$$

which vanishes by virtue of equation (4). Thus, if the Neumann problem has a solution, then necessarily

$$\int_\Omega f d\Omega = 0.$$

On the other hand, if a solution exists it is not unique. Together with u, $u + C$ will also be a solution, where C is an

arbitrary constant. We avoid this arbitrariness by requiring that the solution sought must satisfy the equation

$$\int_{\Omega} u \, d\Omega = 0. \tag{11}$$

We now consider the set of functions which are continuous in $\overline{\Omega}$ together with their first and second derivatives and satisfy equations (4) and (11). We shall prove that the operator L_0 is positive-bounded-below for this set. In fact

$$(L_0 u, u) = -\int_{\Omega} u \sum_{j,k=1}^{m} \frac{\partial}{\partial x_j}\left(A_{jk}\frac{\partial u}{\partial x_j}\right) d\Omega.$$

Integrating by parts and using condition (4), we obtain

$$(L_0 u, u) = \int_{\Omega} \sum_{j,k=1}^{m} A_{jk}\frac{\partial u}{\partial x_j}\frac{\partial u}{\partial x_k} d\Omega \geq \mu_0 \int_{\Omega} \sum_{k=1}^{m}\left(\frac{\partial u}{\partial x_k}\right)^2 d\Omega. \tag{12}$$

The general Poincaré inequality (see § 22)

$$\int_{\Omega} u^2 d\Omega \leq A\int_{\Omega} \sum_{k=1}^{m}\left(\frac{\partial u}{\partial x_k}\right)^2 d\Omega + B\left(\int_{\Omega} u \, d\Omega\right)^2$$

under condition (11) gives

$$\int_{\Omega} \sum_{k=1}^{m}\left(\frac{\partial u}{\partial x_k}\right)^2 d\Omega > \frac{1}{A}\int u^2 d\Omega = \frac{1}{A}\|u\|^2.$$

Substituting this in (12), we obtain

$$(L_0 u, u) \geq \gamma^2 \|u\|^2, \qquad \gamma = \sqrt{\left(\frac{\mu_0}{A}\right)}.$$

§ 25. Degenerate elliptic equations; the Chaplygin equation

We shall call a second-order differential operator

$$-\sum_{j,k=1}^{m} A_{jk}(P)\frac{\partial^2 u}{\partial x_j \partial x_k} + \sum_{j=1}^{m} B_j(P)\frac{\partial u}{\partial x_j} + C(P)u$$

a degenerate elliptic operator in the given finite domain Ω, if the inequality

$$\sum_{j,k=1}^{m} A_{jk}(P)t_j t_k > 0 \tag{1}$$

occurs for all points $P \in \Omega$ and for all values of the real numbers $t_1, t_2, \ldots, t_m$, not simultaneously zero, but the inequality (6) § 24

$$\sum_{j,k=1}^{m} A_{jk}(P)t_j t_k \geq \mu_0 \sum_{j=1}^{m} t_j^2$$

does not hold for all points $P \in \Omega$ and for all values $t_1, t_2, \ldots, t_m$, no matter what the value of the positive constant μ_0.

One of the simplest operators of this kind is the Tricomi operator considered in the previous paragraph.

We shall call a differential equation of the type $Lu = f(P)$, where $f(P)$ is a given function and L is degenerate elliptic operator, *a degenerate elliptic equation*. The general theory of such equations is given in the works of M. V. Keldysh [2], M. I. Vishik [4 – 6], Oleinik [1], N. D. Vvedenskaya [1], the author [13, 15, 16] and in the very interesting book by M. Sh. Birman [4].

A series of important results related to particular forms of degenerate equations is contained in the books of S. A. Chaplygin [1], F. Tricomi [1] and S. Bergman and M. Schiffer [1], these last two containing a bibliography. The closest of all, from the point of view of the methods presented therein, to the present book are the works of M. I. Vishik and of the author. Ye. V. Makhover [1] has considered a degenerate fourth-order equation. Confining ourselves to the case of two independent variables, we consider the finite domain Ω in the half-plane $y > 0$ (Fig. 9) the boundary of which consists of the portion $a \leq x \leq b$ of the x-axis (we shall denote it by S') and of the curve S'' which lies entirely within the upper half-plane except for its ends at $x = a$ and $x = b$. Denoting, as usually, the boundary of the domain Ω by S, we have $S = S' + S''$. Let the function $\varphi(y)$ be defined and, say, continuous over $0 < y < Y$, where Y is not less than the

greatest of the ordinates of the points on the curve S'' and let $\varphi(0)=0$ and $\varphi(y)>0$ for $y>0$. Over the same region $0<y<Y$ there is also a continuous function $\omega(y)$, concerning which we assume that $\omega(y)\geq k$, where k is a positive constant. We consider the elliptic differential operator

$$-\varphi(y)\frac{\partial^2 u}{\partial x^2}-\frac{\partial}{\partial y}\left(\omega(y)\frac{\partial u}{\partial y}\right), \tag{2}$$

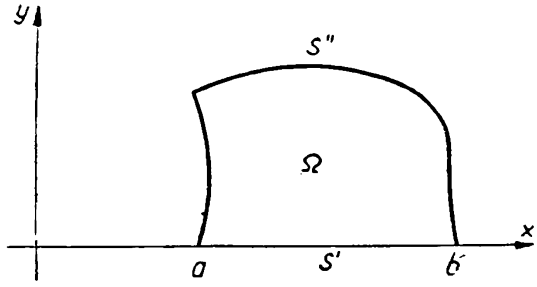

FIG. 9

which is degenerate in the domain Ω; this operator transforms to the Tricomi operator for $\varphi(y)\equiv y$, $\omega(y)\equiv 1$. We pose the problem of integrating the equation

$$-\varphi(y)\frac{\partial^2 u}{\partial x^2}-\frac{\partial}{\partial y}\left(\omega(y)\frac{\partial u}{\partial y}\right)=f(x, y), \tag{3}$$

in the domain Ω, where $f(x, y)$ is a given function, under one of the two types of homogeneous boundary conditions

I $$u|_S=0, \tag{4}$$

II $$u|_{S'}=0, \quad \left.\frac{\partial u}{\partial y}\right|_{S'}=0. \tag{5}$$

We shall denote operator (2), considered for the set of functions which are twice continuously differentiable in $\overline{\Omega}=\Omega+S$ and satisfy boundary conditions (4) or (5), by L_{I} or L_{II} respectively. We shall prove that *the operators* L_{I} *and* L_{II} *are positive-bounded-below.*

Let $u(x, y)$ and $v(x, y)$ be two functions which belong simultaneously either to the field of definition D_{L_I} of operator L_I or to the field of definition $D_{L_{II}}$ of operator L_{II}. In particular, both functions satisfy simultaneously either boundary condition (4) or boundary conditions (5). Understanding by i either of the indices I or II, we construct the scalar product

$$(L_i u, v) = -\int_\Omega\int v\left[\varphi(y)\frac{\partial^2 u}{\partial x^2}+\frac{\partial}{\partial y}\left(\omega(y)\frac{\partial u}{\partial y}\right)\right]\mathrm{d}x\mathrm{d}y.$$

Applying Green's formula to the latter integral, we find

$$(L_i u, v)=\int_\Omega\int\left[\varphi(y)\frac{\partial u}{\partial x}\frac{\partial v}{\partial x}+\omega(y)\frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\right]\mathrm{d}x\mathrm{d}y$$

$$-\int_S v\left[\varphi(y)\frac{\partial u}{\partial x}\cos(n, x)+\omega(y)\frac{\partial u}{\partial y}\cos(n, y)\right]\mathrm{d}S; \qquad (6)$$

n, as usual, denotes the outward drawn normal to S. It is easily seen that the line integral equals zero. In fact, both under condition (4) and under condition (5) the function v vanishes on the curve S'' and it suffices to take the line integral in (6) along the portion S' of the x-axis. If condition (4) is given, $v|_{s'}=0$ and the integral along S' equals zero. If condition (5) is given, $\dfrac{\partial u}{\partial y}=0$ and moreover, $\cos(n, x)=0$ and, as before, the integral along S' vanishes. Now

$$(L_i u, v)=\int_\Omega\int\left[\varphi(y)\frac{\partial u}{\partial x}\frac{\partial v}{\partial x}+\omega(y)\frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\right]\mathrm{d}x\mathrm{d}y; \quad i=\mathrm{I},\ \mathrm{II}. \qquad (7)$$

Formula (7) shows that the operators L_I and L_{II} are symmetric, since the right-hand side of the formula mentioned is unchanged under the permutation of the functions u and v. Putting $u=v$ in (7) we find

$$(Lu_i, u)=\int_\Omega\int\left[\varphi(y)\left(\frac{\partial u}{\partial x}\right)^2+\omega(y)\left(\frac{\partial u}{\partial y}\right)^2\right]\mathrm{d}x\mathrm{d}y; \quad i=\mathrm{I},\ \mathrm{II}. \qquad (8)$$

From formula (8) it immediately follows that the operators L_{I} and L_{II} are positive definite. In order to prove that they are positive-bounded-below we enclose Ω in a rectangle (Fig. 10) defined by the inequalities $a' \leq x \leq b'$ $0 \leq y \leq Y$; obviously

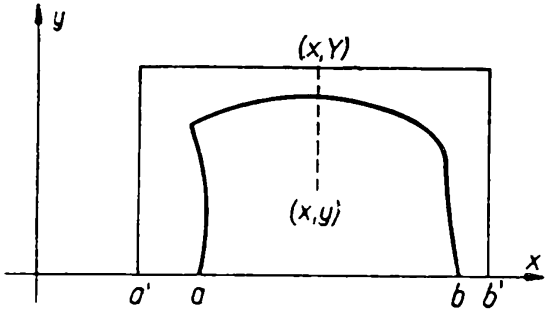

FIG. 10

$a' \leq a$ and $b' \geq b$. We denote this rectangle by the symbol Π. Let there be given a function $u(x, y)$ from the field of definition of the operator L_{I} or the operator L_{II}, i.e. twice continuously differentiable in $\bar{\Omega}$ and satisfying on S either condition (4) or conditions (5). We define the function $u(x, y)$ in the remainder of the rectangle Π by assuming it to equal zero outside Ω. With such a definition this function will be continuous in the rectangle and its first derivatives will be discontinuous only at the curve S'' where there will be a finite jump. Let (x, y) be a point within the rectangle. We have

$$\int_y^Y \frac{\partial u(x, \eta)}{\partial \eta} \mathrm{d}\eta = u(x, Y) - u(x, y).$$

But the point (x, Y) lies outside Ω, and thus at this point the function u becomes zero. Hence

$$u(x, y) = -\int_y^Y \frac{\partial u(x, \eta)}{\partial \eta} \mathrm{d}\eta.$$

By Cauchy's inequality

$$u^2(x, y) \leq (Y-y)\int_y^Y \left(\frac{\partial u(x, \eta)}{\partial \eta}\right)^2 d\eta \leq Y\int_0^Y \left(\frac{\partial u}{\partial y}\right)^2 dy$$

$$\leq \frac{Y}{k}\int_0^Y \left[\varphi(y)\left(\frac{\partial u}{\partial x}\right)^2 + \omega(y)\left(\frac{\partial u}{\partial y}\right)^2\right] dy.$$

Integrating this inequality over the rectangle Π, we obtain

$$\int_\Pi\int u^2 dxdy \leq \frac{Y^2}{k}\int_\Pi\int \left[\varphi(y)\left(\frac{\partial u}{\partial x}\right)^2 + \omega(y)\left(\frac{\partial u}{\partial y}\right)^2\right] dxdy.$$

In the domain $\Pi - \Omega$ we have $u(x, y) \equiv 0$; dropping the integral over this region which equals zero, we arrive at the inequality

$$\int_\Omega\int u^2 d\Omega \leq \frac{Y^2}{k}\int_\Omega\int \left[\varphi(y)\left(\frac{\partial u}{\partial x}\right)^2 + \omega(y)\left(\frac{\partial u}{\partial y}\right)^2\right] dxdy. \qquad (9)$$

The comparison of formulae (8) and (9) gives

$$(L_i u, u) \geq \frac{k}{Y^2}\int_\Omega\int u^2 dxdy = \frac{k}{Y^2}\|u\|^2; \quad i = \mathrm{I}, \mathrm{II},$$

and this also means that the operators L_{I} and L_{II} are positive-bounded-below.

Using the results of §§ 11, 18 and 19, we can make the following assertions:

1) A solution (at least a generalized solution) of the degenerate elliptic equation (3) under boundary conditions (4) and (5) exists and can be obtained as the solution of the problem of the minimum of the functional

$$F(u) = \int_\Omega\int \left[\varphi(y)\left(\frac{\partial u}{\partial x}\right)^2 + \omega(y)\left(\frac{\partial u}{\partial y}\right)^2 - 2uf\right] dxdy \qquad (10)$$

with respect to the sets of functions such that either $u|_S = 0$ if condition (4) is given, or $u|_{S'} = 0$ if condition (5) is given. In the latter case the condition $\left.\frac{\partial u}{\partial y}\right|_{S'} = 0$ is natural.

2) The solution of the homogeneous equation

$$\varphi(y)\frac{\partial^2 u}{\partial x^2}+\frac{\partial}{\partial y}\left(\omega(y)\frac{\partial u}{\partial y}\right)=0 \tag{11}$$

under the non-homogeneous boundary conditions

$$u|_S=g_1(s), \tag{12}$$

where s is the length of arc of the contour S, reduces to finding the minimum of the functional

$$\int_\Omega\int\left[\varphi(y)\left(\frac{\partial u}{\partial x}\right)^2+\omega(y)\left(\frac{\partial u}{\partial y}\right)^2\right]\mathrm{d}x\mathrm{d}y \tag{13}$$

with respect to the set of functions satisfying condition (12) provided that there exists at least one function which satisfies condition (12) and imparts a finite value to the integral (13).

3) The solution of equation (11) under the boundary conditions

$$u|_{S'}=g_1(s),\quad \left.\frac{\partial u}{\partial y}\right|_{S'}=g_2(x) \tag{14}$$

reduces to finding the minimum of the functional

$$\int_\Omega\int\left[\varphi(y)\left(\frac{\partial u}{\partial x}\right)^2+\omega(y)\left(\frac{\partial u}{\partial y}\right)^2\right]\mathrm{d}x\mathrm{d}y-2\int_a^b u(x)g_2(x)\,\mathrm{d}x \tag{15}$$

with respect to the set of functions satisfying the condition $u|_{S'}=g_1(x)$ provided that there exists at least one function which satisfies this condition and imparts a finite value to functional (14). For this functional the condition

$$\left.\frac{\partial u}{\partial y}\right|_{S'}=g_2(x)$$

is natural.

A particularly important special case of equation (11) is the Chaplygin equation [1];

$$\frac{1-(2\beta+1)\tau}{2\tau(1-\tau)^{\beta+1}}\frac{\partial^2\psi}{\partial\theta^2}+\frac{\partial}{\partial\tau}\left[\frac{2\tau}{(1-\tau)^\beta}\frac{\partial\psi}{\partial\tau}\right]=0. \tag{16}$$

Here β is a positive constant connected by the relationship $\beta=(\gamma-1)^{-1}$ with the ratio γ of the specific heats. In the sense of the problem the quantity τ is strictly positive; to values of $\tau<(2\beta+1)^{-1}$ there correspond velocities of the gas particles less than the velocity of sound; for $\tau=(2\beta+1)^{-1}$ the velocity of the corresponding gas particles coincides with the velocity of sound. The Chaplygin equation is of elliptic type in the region of subsonic velocities; it degenerates in the (θ, τ) plane on the straight line $\tau=(2\beta+1)^{-1}$, i.e. where the velocities of gas particles reach the velocity of sound. Equation (16) is transformed to form (11) if we put

$$x=\theta, \quad y=\frac{1}{2\beta+1}-\tau.$$

From what has been said above, it follows that the Chaplygin equation can be solved using the energy method if the gas flow is subsonic (but allowing the attainment of the velocity of sound), and if the domain Ω of the variation of the variables θ and τ and, say, the value of the unknown function ψ on the contour of this region are given. As usual, we must assume for this that there exists a function which assumes the specified values at the contour of the domain Ω and which gives a finite value to the integral

$$\int_{\Omega}\int \left[\frac{1-(2\beta+1)\tau}{2\tau(1-\tau)^{\beta+1}}\left(\frac{\partial\psi}{\partial\theta}\right)^2+\frac{2\tau}{(1-\tau)^{\beta}}\left(\frac{\partial\psi}{\partial\tau}\right)^2\right]\mathrm{d}\theta\,\mathrm{d}\tau.$$

We now turn to operator (2). The picture becomes somewhat more complex if the degeneracy is occasioned by the vanishing of the function $\omega(y)$. Without going into details, for which we refer the reader to the works of M. I. Vishik and of the author cited above, we merely note here the basic results. As before, we assume that the domain has the form depicted in Fig. 9. We suppose that for $0\leq y\leq Y$ the function $\varphi(y)\geq 0$, and $\omega(y)=y^{\alpha}\omega_1(y)$, $\omega_1(y)\geq k$, α and k being positive constants. If $\alpha<1$ we can specify the value of the function sought over the entire contour S and in this case the operator

(2) is positive-bounded-below and, consequently, the corresponding boundary value problem has a unique solution, which can be found with the aid of the energy method. If $\alpha \geq 1$, it is impossible to specify anything for the degeneracy line S'. In this case, if $1 \leq \alpha \leq 2$, operator (2) is positive-bounded-below for the set of functions equal to zero on S'' and the problem of integrating equation (3) under the boundary condition

$$u|_{S''} = 0 \tag{17}$$

always has a solution which is unique; if $\alpha > 2$, operator (2) under boundary condition (17) is merely positive definite but not positive-bounded-below, and the problem of integrating equation (3) under boundary condition (17) is insoluble in the general case, if we require that the integral sought possess finite energy.

§ 26. The principle of minimum of potential energy in elasticity theory

We shall use the following notation: $\mathbf{u}$ (u_x, u_y, u_z) is the displacement vector, $\sigma_x, \ldots, \tau_{xy}, \ldots \tau_{yz}$ are the stress components, $\mathbf{K}$ (X, Y, Z) is the force per unit volume. Further we shall denote by ε_x the relative extension in the direction of the x-axis and by γ_{xy} the shear in the xy-plane, $\varepsilon_y, \varepsilon_z, \gamma_{yz}, \gamma_{xz}$ being analogously defined. We shall call the quantities $\varepsilon_x, \varepsilon_y, \varepsilon_z, \gamma_{xy}, \gamma_{yz}, \gamma_{xz}$, the deformation components, although this is not entirely correct.

They are connected with the displacement vector $\mathbf{u}$ by the relations

$$\varepsilon_x = \frac{\partial u_x}{\partial x}, \quad \gamma_{xy} = \frac{\partial u_x}{\partial y} + \frac{\partial u_y}{\partial x} \tag{1}$$

and by four others analogous to them. In an elastic body the components of stress and deformation are related by the generalized Hooke's law

$$\left.\begin{aligned}
\sigma_x &= a_{11}\varepsilon_x + a_{12}\varepsilon_y + a_{13}\varepsilon_z + a_{14}\gamma_{xy} + a_{15}\gamma_{xz} + a_{16}\gamma_{yz},\\
\sigma_y &= a_{21}\varepsilon_x + a_{22}\varepsilon_y + a_{23}\varepsilon_z + a_{24}\gamma_{xy} + a_{25}\gamma_{xz} + a_{26}\gamma_{yz},\\
\sigma_z &= a_{31}\varepsilon_x + a_{32}\varepsilon_y + a_{33}\varepsilon_z + a_{34}\gamma_{xy} + a_{35}\gamma_{xz} + a_{36}\gamma_{yz},\\
\tau_{xy} &= a_{41}\varepsilon_x + a_{42}\varepsilon_y + a_{43}\varepsilon_z + a_{44}\gamma_{xy} + a_{45}\gamma_{xz} + a_{46}\gamma_{yz},\\
\tau_{xz} &= a_{51}\varepsilon_x + a_{52}\varepsilon_y + a_{53}\varepsilon_z + a_{54}\gamma_{xy} + a_{55}\gamma_{xz} + a_{56}\gamma_{yz},\\
\tau_{yz} &= a_{61}\varepsilon_x + a_{62}\varepsilon_y + a_{63}\varepsilon_z + a_{64}\gamma_{xy} + a_{65}\gamma_{xz} + a_{66}\gamma_{yz}.
\end{aligned}\right\} \tag{2}$$

The quantities a_{jk} do not depend upon the deformed state of the body and are mechanical characteristics of the material. They are known as the *elastic constants* of the given material.

It should, however, be noted that the a_{jk} will be constant only for a homogeneous body. In the general case of a non-homogeneous body, they will be functions of the coordinates x, y, z.

The quantities a_{jk} are related by the important relationship

$$a_{jk} = a_{kj}, \tag{3}$$

This follows from the existence of an elastic potential, i.e. a function of the deformation components $W(\varepsilon_x, \varepsilon_y, \ldots, \gamma_{yz})$ such that

$$\sigma_x = \frac{\partial W}{\partial \varepsilon_x}, \quad \sigma_y = \frac{\partial W}{\partial \varepsilon_y}, \quad \ldots \quad \tau_{yz} = \frac{\partial W}{\partial \gamma_{yz}}.$$

From formulae (2) and (3) it readily follows that

$$2W = \varepsilon_x\sigma_x + \varepsilon_y\sigma_y + \varepsilon_z\sigma_z + \gamma_{xy}\tau_{xy} + \gamma_{xz}\tau_{xz} + \gamma_{yz}\tau_{yz}. \tag{4}$$

In the sequel we shall denote the elastic potential by $W(\mathbf{u})$. Substituting the stress components according to formulae (2), we find that W is a quadratic form (see e.g. I. M. Gel'fand [1]) in the deformation components. In elasticity theory it is proved that this form is positive-bounded-below, i.e., that $W \geq 0$, W equalling zero when and only when all the deformation components equal zero (see, e. g. L. S. Leibenzon [2]). Equation (2) is valid for the very general case of anisotropic media. If the elastic medium is isotropic then the equations

(2) simplify considerably and reduce to the well-known *Lamé equations:*

$$\sigma_x=\lambda\theta+2\mu\varepsilon_x, \quad \tau_{xy}=\mu\gamma_{xy},$$
$$\sigma_y=\lambda\theta+2\mu\varepsilon_y, \quad \tau_{xz}=\mu\gamma_{xz},$$
$$\sigma_z=\lambda\theta+2\mu\varepsilon_z, \quad \tau_{yz}=\mu\gamma_{yz},$$
$$\theta=\operatorname{div}\mathbf{u}.$$

Here λ and μ are the so called Lamé constants.

The well-known equilibrium equations which we write in the rather unusual form

$$\left.\begin{aligned}-\left(\frac{\partial\sigma_x}{\partial x}+\frac{\partial\tau_{xy}}{\partial y}+\frac{\partial\tau_{xz}}{\partial z}\right)&=X\\ -\left(\frac{\partial\tau_{xy}}{\partial x}+\frac{\partial\sigma_y}{\partial y}+\frac{\partial\tau_{yz}}{\partial z}\right)&=Y\\ -\left(\frac{\partial\tau_{xz}}{\partial x}+\frac{\partial\tau_{yz}}{\partial y}+\frac{\partial\sigma_z}{\partial z}\right)&=Z\end{aligned}\right\} \tag{5}$$

can be considered as a single vectorial equation in terms of the displacement vector $\mathbf{u}$. For this it suffices to substitute in the left-hand sides of equation (5), considered as the components of some vector, the expressions (2) in place of σ_x, τ_{xy}, ... τ_{yz} and further to substitute the deformation components given by formula (1). We denote the vector quantity mentioned by $\mathbf{Au}$, its components being $A_x\mathbf{u}$, $A_y\mathbf{u}$, $A_z\mathbf{u}$. System (5) is now written in the form of a single vector equation

$$\mathbf{Au}=\mathbf{K}. \tag{6}$$

We note that in the most important case of an isotropic medium the operator $\mathbf{A}$ has the extremely simple form:

$$\mathbf{Au}=(\lambda+2\mu)\nabla(\nabla\cdot\mathbf{u})-\mu\nabla\times(\nabla\times\mathbf{u}).$$

Let $\mathbf{u}'$ and $\mathbf{u}''$ be the two vectors of the elastic displacement corresponding to the forces per unit volume $\mathbf{K}'$ and $\mathbf{K}''$. We recall the well-known Betti formulae (see e. g. E. Trefftz [2])

$$\int_\Omega \mathbf{u}'\mathbf{Au}''\,d\Omega=2\int_\Omega W(\mathbf{u}',\mathbf{u}'')d\Omega-\int_S \mathbf{u}'\mathbf{t}''^{(n)}dS, \tag{7}$$

$$\int_{\Omega} \mathbf{u}\mathbf{A}\mathbf{u}\,d\Omega = 2\int_{\Omega} W(\mathbf{u})\,d\Omega - \int_{S} \mathbf{u}\mathbf{t}^{(n)}\,dS, \tag{8}$$

$$\int_{\Omega} \mathbf{u}'(\mathbf{A}\mathbf{u}'' - \mathbf{u}''\mathbf{A}\mathbf{u}')\,d\Omega = -\int_{S} (\mathbf{u}'\mathbf{t}''^{(n)} - \mathbf{u}''\mathbf{t}'^{(n)})\,dS. \tag{9}$$

Here we use the following notation:

1) One dash (two dashes) denotes a quantity connected with the vector $\mathbf{u}'$ (the vector $\mathbf{u}''$);

2) $\mathbf{t}^{(n)}$ is the stress vector acting on the surface of the body;

3) $W(\mathbf{u}', \mathbf{u}') = \frac{1}{2}(\varepsilon_x'\sigma_x'' + \varepsilon_y'\sigma_y'' + \varepsilon_z'\sigma_z'' + \gamma_{xy}'\tau_{xy}'' + \gamma_{xz}'\tau_{xz}'' + \gamma_{yz}'\tau_{yz}'')$; as is known $W(\mathbf{u}', \mathbf{u}'') = W(\mathbf{u}'', \mathbf{u}')$,

We now turn to the problem of integrating the equations of elasticity theory. We shall consider the boundary conditions accompanying these equations to be homogeneous.

The simplest and most important types of boundary conditions in elasticity theory are the following:

(a) the boundary S is fixed:

$$\mathbf{u}|_S = 0; \tag{10}$$

(b) the boundary of the body is free from the influence of external forces:

$$\mathbf{t}^{(n)}|_S = 0; \tag{11}$$

(c) the boundary of the body consists of two parts S_1 and S_2 of which the first is fixed and the second free from influence of external forces:

$$\mathbf{u}|_{S_1} = 0, \quad \mathbf{t}^{(n)}|_{S_2} = 0. \tag{12}$$

Problem (c) is often known as the mixed problem in elasticity theory.

We shall prove that all the problems enumerated are equivalent to the problem of finding the minimum of the integral

$$\int_{\Omega} \{W(\mathbf{u}) - \mathbf{u}\mathbf{K}\}\,d\Omega \tag{13}$$

under certain conditions. This statement is the well-known principle of minimum potential energy in elasticity theory,

since integral (13) equals the potential energy of an elastic body subject to the influence of volume forces $\mathbf{K}$.

We construct the expression

$$(\mathbf{Au}, \mathbf{u}) = \int_{\Omega} \mathbf{u} \cdot \mathbf{Au} d\Omega.$$

We apply formula (8) to this integral and note that under either of conditions (10) or (12) the integrand function in the surface integral vanishes. Thus

$$(\mathbf{Au}, \mathbf{u}) = 2\int_{\Omega} W(\mathbf{u}) d\Omega \geq 0. \tag{14}$$

Further, if $(\mathbf{Au}, \mathbf{u}) = 0$, then $W(\mathbf{u}) \equiv 0$, and since W is a positive-bounded-below quadratic form in the deformation components, the latter is identically equal to zero. But then $\mathbf{u}$ is the vector of a small rigid displacement of the body; if the boundary of the body (condition [10]) or part of the boundary (condition [12]) is fixed then a rigid displacement is impossible and $\mathbf{u} \equiv 0$. Thus, the operator $\mathbf{A}$ is positive definite for each of the sets defined by conditions (10) or (12). By the minimum functional theorem (§ 11), finding the integral of equation (6) (or, what is the same thing, of the system of equations of elasticity theory) under one of conditions (10) or (12) is equivalent to finding a vector which satisfies the same boundary conditions and minimizes the functional

$$F(\mathbf{u}) = (\mathbf{Au}, \mathbf{u}) - 2(\mathbf{K}, \mathbf{u}) = 2\int_{\Omega} \{W(\mathbf{u}) - \mathbf{Ku}\} d\Omega.$$

Thus, the principle of minimum potential energy is proved as applied to problems (a) and (c).

We now turn to problem (b), the role of which in elasticity theory is the same as the role of the Neumann problem in the theory of harmonic functions. For the linear set of vectors which are continuous in $\bar{\Omega}$ together with their first and second derivatives and satisfy boundary condition (11) the operator $\mathbf{A}$ will not be positive definite. In fact formula (14) still holds

true, so that $(\mathbf{Au}, \mathbf{u}) \geq 0$. However $(\mathbf{Au}, \mathbf{u})$ can equal zero not only for $\mathbf{u} \equiv 0$ – it suffices to take for $\mathbf{u}$ an arbitrary vector of a small rigid displacement. As in the Neumann problem we make $\mathbf{A}$ positive definite by considering it for a more restricted set of vectors.

Problem (b) does not always have a solution and it is easy to find the necessary conditions for its solubility.

They consist of the fact that the totality of the forces applied to the elastic body must be statically equivalent to zero. Since surface forces are absent [condition (11)], the forces per unit volume must be statically equivalent to zero:

$$\int_{\Omega} \mathbf{K} d\Omega = 0; \qquad \int_{\Omega} \mathbf{r} \times \mathbf{K} d\Omega = 0.$$

Here $\mathbf{r}$ is the radius vector of a variable point in the region Ω and the cross denotes vector multiplication of the vectors. Further, by condition (11) the displacement vector is not defined uniquely – two elastic displacement vectors corresponding to the same volume forces and the same condition (11) can differ by a vector corresponding to a small arbitrary rigid displacement. This arbitrary vector can be fixed by subjecting, for example, the solution sought to the conditions

$$\int_{\Omega} \mathbf{u} d\Omega = 0; \qquad \int_{\Omega} \mathbf{r} \times \mathbf{u} d\Omega = 0. \tag{15}$$

In order to be convinced of this it suffices to prove that the vector associated with the small rigid displacement satisfying equations (15) equals zero. In fact the components $u_x^{(0)}$, $u_y^{(0)}$, $u_z^{(0)}$ of a small rigid displacement have the form

$$u_x^{(0)} = a + \alpha y - \beta z,$$

$$u_y^{(0)} = b - \alpha x + \gamma z,$$

$$u_z^{(0)} = c + \beta x - \gamma y,$$

where, a, b, c, α, β, γ are constants. As is known, if the vector satisfies equation (15) in some coordinate system, then it also

satisfies it in any other system. We therefore adopt as the coordinate axes the principal axes of inertia of the body Ω at its centroid. With such a choice the static moments equal zero and the first of equations (15) immediately gives $a=b=c=0$.

Further, taking the x-component of the second of equations (15) we obtain

$$0=\int_{\Omega}(yu_z^{(0)}-zu_y^{(0)})\mathrm{d}\Omega=\beta I_{xy}-\gamma I_{yy}-\alpha I_{xz}-\gamma I_{zz}.$$

The products of inertia I_{xy}, I_{xz} equal zero and from the latter equality it follows that $\gamma=0$. Similarly it can be proved that $\alpha=\beta=0$.

We now consider the linear set of displacement vectors which satisfy not only condition (11) but also conditions (15). It is easily seen that for this linear set the operator $\mathbf{A}$ is positive definite. It is sufficient to prove that this time it follows from the equality $(\mathbf{Au}, \mathbf{u})=0$ that $\mathbf{u}=0$. But, as was stated above it follows from this equation that $\mathbf{u}$ is the vector of a small rigid displacement and, as this vector satisfies equation (15), it equals zero.

Thus the principle of minimum potential energy remains true for problem (b) if we assume that both the solution of the differential equations of elasticity theory, and the vector minimizing functional (13) satisfy equations (15).

Boundary condition (11) is natural and it is not necessary to satisfy it beforehand in seeking the minimum.

By reasoning as in § 18 it is also possible to find the form of the minimum integral in the case of non-homogeneous boundary conditions readily.

In the general case of the mixed problem this integral has the form

$$\int_{\Omega}\{W(\mathbf{u})-\mathbf{uK}\}\mathrm{d}\Omega-\int_{S_1}\mathbf{ut}^{(n)}\mathrm{d}S. \tag{16}$$

Here $\mathbf{t}^{(n)}$ is the stress vector specified on the part S_2 of the surface S. In particular, if the displacements are specified for the entire surface, the minimum integral assumes the form

$$\int_{\Omega}\{W(\mathbf{u})-\mathbf{uK}\}\mathrm{d}\Omega. \tag{17}$$

If the stress vector is specified for the entire surface, integral (16) is transformed to the following

$$\int_{\Omega}\{W(\mathbf{u})-\mathbf{uK}\}\mathrm{d}\Omega-\int_{S}\mathbf{ut}^{(n)}\mathrm{d}S. \tag{18}$$

The boundary conditions should be satisfied beforehand only for that part of the surface upon which the displacements are specified.

It has been established above that the operator $\mathbf{A}$ which enters into the equations of elasticity theory, is positive definite for the set of vector displacements satisfying conditions (10), (12) or (11) and (15). It can be proved that for each of these sets the operator $\mathbf{A}$ is positive-bounded-below. The reader will find the detailed formulation and proof of the corresponding theorems and also a bibliographic survey in the author's book [11]. The fact that the operator is positive-bounded-below will be proved below (§ 39) for the case of a fixed boundary.

§ 27. The bending of thin plates

1. In the previous section of this chapter we have started from the differential equations and the boundary conditions of a given problem and reduced them to the problem of the minimum of a certain functional. In the present paragraph we shall show how it is possible to obtain the differential equation and the natural boundary conditions starting from the variational problem.

As is known, it is possible to construct the theory of thin

plates by utilizing the hypotheses formulated below. (A rather different formulation is given by Kirchhoff [1].)

As usual we let the xy-plane be the central plane of the plate and denote the thickness of the plate by h.

HYPOTHESES: 1) points in the central plane of the disc undergo only normal displacements; 2) in view of the small thickness of the plate the displacements u_x and u_y can be considered to be linear functions of z, and the displacement u_z to be independent of z; 3) the shears in the vertical plane equal zero at points in the central plane; 4) in comparison with σ_z, σ_y, τ_{xy} the stresses τ_{xz}, τ_{yz}, σ_z are negligibly small; 5) the principle of minimum potential energy still holds true for the bent plate in a form to which we shall refer below.

We denote by $w(x, y)$ the normal displacements of points (x, y) of the central plane. In virtue of hypotheses 1) and 2) $u_x=z\tilde{u}\ (x, y)$, $u_y=z\tilde{v}\ (x, y)$, where $\tilde{u}$, $\tilde{v}$ are certain functions of x and y. By hypothesis 3) we have $\gamma_{xz}=\gamma_{yz}=0$ for $z=0$; hence $\tilde{u}=-\dfrac{\partial w}{\partial x}$, $\tilde{v}=-\dfrac{\partial w}{\partial y}$ and consequently,

$$u_x=-z\frac{\partial w}{\partial x},\quad u_y=-z\frac{\partial w}{\partial y}. \tag{1}$$

Using hypothesis 4), we put $\sigma_z\equiv 0$. From the Lamé equations we then find $\lambda\theta+2\mu\varepsilon_z=0$, whence

$$\varepsilon_z=-\frac{\lambda}{\lambda+2\mu}(\varepsilon_x+\varepsilon_y).$$

Substituting this in the first two Lamé formulae and using the well-known relationships between the elastic constants of an isotropic body, we find

$$\sigma_x=\frac{E}{1-\sigma^2}(\varepsilon_x+\sigma\varepsilon_y),\quad \sigma_y=\frac{E}{1-\sigma^2}(\varepsilon_y+\sigma\varepsilon_x), \tag{2}$$

where E is Young's modulus and σ Poisson's ratio for the plate material. We further note the Lamé equation connecting τ_{xy} and γ_{xy}:

$$\tau_{xy}=\mu\gamma_{xy}=\frac{E}{2(1+\sigma)}\gamma_{xy}. \tag{3}$$

We denote by Ω the spatial domain occupied by the plate, by S and S_1 its upper and lower bases and by $\sum$ its lateral surface. As usual we assume that a normal load of intensity $q(x, y)$ acts upon the upper base S and that the lower base S_1 is free from the influence of external forces. We neglect the influence of volume forces on the plate. If the edge of the plate $\sum$ is rigidly fixed, then, corresponding to formula (16) of the previous paragraph, we conclude that the problem of the bending of the plate reduces to the problem of the minimum of the functional

$$\int\int_{\Omega}\int W(\mathbf{u})\,d\Omega - \int_{S}\int q u_z\,dS \tag{4}$$

under the corresponding boundary condition. We assume that *form* (4) *of the minimal functional still holds in the case when part of the edge or the whole edge of the plate is free or rests freely*. Further, with the aim of simplifying the exposition, we shall restrict ourselves to the case where part of the boundary is rigidly fixed and the other part rests freely.

We now simplify functional (4). We have

$$W(\mathbf{u}) = \frac{1}{2}(\sigma_x\varepsilon_x + \sigma_y\varepsilon_y + \tau_{xy}\gamma_{xy} + \sigma_z\varepsilon_z + \tau_{xz}\gamma_{xz} + \tau_{yz}\gamma_{yz}).$$

Hypothesis 4) allows us to neglect the last three terms in comparison with the first three; further, using formulae (1)–(3), we obtain

$$W(\mathbf{u}) = \frac{Ez^2}{2(1-\sigma^2)}\left\{(\nabla^2 w)^2 + 2(1-\sigma)\left[\left(\frac{\partial^2 w}{\partial x \partial y}\right)^2 - \frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}\right]\right\}.$$

Hence

$$\int\int_{\Omega}\int W(\mathbf{u})\,d\Omega = \int_{S}\int dS \int_{-\frac{h}{2}}^{\frac{h}{2}} W(\mathbf{u})\,dz$$

$$= \frac{D}{2}\int_{S}\int\left\{(\nabla^2 w)^2 + 2(1-\sigma)\left[\left(\frac{\partial^2 w}{\partial x \partial y}\right)^2 - \frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}\right]\right\} dS; \tag{5}$$

$$D=\frac{Eh^3}{12(1-\sigma^2)}.$$

In what follows it is important that the integral

$$\int_S\int\left[\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2-\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}\right]\mathrm{d}S$$

can be transformed into an integral taken along the contour of the region S; we denote this contour by L. Integrating by parts we find

$$\int_S\int\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2\mathrm{d}S=-\int_S\int\frac{\partial w}{\partial x}\frac{\partial^3 w}{\partial x\partial y^2}\,\mathrm{d}S+\int_L\frac{\partial w}{\partial x}\frac{\partial^2 w}{\partial x\partial y}\,\mathrm{d}x,$$

$$\int_S\int\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}\mathrm{d}S=-\int_S\int\frac{\partial w}{\partial x}\frac{\partial^3 w}{\partial x\partial y^2}\,\mathrm{d}S+\int_L\frac{\partial w}{\partial x}\frac{\partial^2 w}{\partial y^2}\,\mathrm{d}y.$$

By subtracting the latter equations we find

$$\int_S\int\left[\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2-\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}\right]\mathrm{d}S=\int_L\frac{\partial w}{\partial x}\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\partial w}{\partial y}\right)\mathrm{d}s;\qquad(6)$$

where $\mathrm{d}s$ is an element of length of the contour L.

We take the line integral by parts. Since the contour L is closed and the function w and its derivatives are obviously of constant sign, the extra integral term drops out and we obtain

$$\int_S\int\left[\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2-\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}\right]\mathrm{d}S=-\int_L\frac{\partial w}{\partial y}\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\partial w}{\partial x}\right)\mathrm{d}s.$$

From the last two equations it follows that

$$\int_S\int\left[\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2-\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}\right]\mathrm{d}S$$

$$=-\frac{1}{2}\int_L\left[\frac{\partial w}{\partial y}\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\partial w}{\partial x}\right)-\frac{\partial w}{\partial x}\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\partial w}{\partial y}\right)\right]\mathrm{d}s.\qquad(6_1)$$

We note that if the function $w(x, y)$ satisfies conditions (8) (see below) on the contour L, then

$$\int\int_S \left[\left(\frac{\partial^2 w}{\partial x \partial y}\right)^2 - \frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}\right] dS = 0. \tag{6_2}$$

By virtue of hypothesis 2) we can put $u_z = w(x, y)$ in the second integral in (4). Combining formulae (5) and (6), we obtain for the potential energy of the bent plate the expression

$$\frac{D}{2}\int\int_S \left[(\nabla^2 w)^2 - \frac{2q}{D}w\right] dS$$
$$+ \frac{D(1-\sigma)}{2}\int_L \left[\frac{\partial w}{\partial y}\frac{d}{ds}\left(\frac{\partial w}{\partial x}\right) - \frac{\partial w}{\partial x}\frac{d}{ds}\left(\frac{\partial w}{\partial y}\right)\right] ds. \tag{7}$$

We now elucidate the boundary conditions to which the function should be subjected. If the lateral surface of the plate is rigidly fixed then at this surface

$$u_x = u_y = u_z = 0.$$

Bearing in mind formula (1) and also the identity $u_z \equiv w$, pointed out above, we find that the conditions for the rigid fixing of the edge of the plate have the form

$$w = 0, \quad \frac{\partial w}{\partial x} = \frac{\partial w}{\partial y} = 0,$$

which may be replaced by the two equivalent equations

$$w = 0, \quad \frac{\partial w}{\partial n} = 0, \tag{8}$$

which are necessarily satisfied for the rigid fixing of part of the contour L. For the freely supported part of the contour L, one of the boundary conditions has the form

$$w = 0. \tag{9}$$

The other condition we obtain as a natural one for functional (7). We further note, without going into details, that

for the free part of the contour two natural conditions are obtained.

We now derive the equilibrium equation for the bent plate. We suppose that a part L_1 of the contour L is rigidly fixed and the remainder L_2 rests freely. Then over the entire contour L the condition $w=0$ is satisfied, moreover $\frac{\partial w}{\partial n}=0$ on L_1.

Let the function $w(x, y)$ minimize functional (7) [which for brevity we denote by $\Psi(w)$] and let η be an arbitrary function which is differentiable a sufficient number of times and satisfies conditions (8) and (9) and let t be an arbitrary real number. The function $w+t\eta$ satisfies the same boundary conditions (8) and (9), therefore

$$\Psi(w+t\eta)\geq\Psi(w).$$

If the function $\eta(x, y)$ is fixed, $\Psi(w+t\eta)$ is a function of t which, as the last inequality shows, has a minimum at $t=0$, and so

$$\frac{\mathrm{d}}{\mathrm{d}t}\Psi(w+t\eta)|_{t=0}=0;$$

which reduces by trivial manipulation to

$$\int_S\int\left(\nabla^2 w\,\nabla^2\eta-\frac{q}{D}\eta\right)\mathrm{d}S+\frac{1-\sigma}{2}\int_L\left\{\frac{\partial\eta}{\partial y}\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\partial w}{\partial x}\right)+\frac{\partial w}{\partial x}\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\partial\eta}{\partial x}\right)\right.$$
$$\left.-\frac{\partial\eta}{\partial x}\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\partial w}{\partial y}\right)-\frac{\partial w}{\partial x}\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\partial\eta}{\partial y}\right)\right\}\mathrm{d}s=0.$$

Further, by Green's formula

$$\int_S\int\nabla^2 w\nabla^2\eta\,\mathrm{d}S=\int_S\int\eta\nabla^4 w\,\mathrm{d}S+\int_L\left(\nabla^2 w\frac{\partial\eta}{\partial n}-\eta\frac{\partial\nabla^2 w}{\partial n}\right)\mathrm{d}s$$
$$=\int_S\int\eta\nabla^4 w\,\mathrm{d}S+\int_L\nabla^2 w\frac{\partial\eta}{\partial n}\mathrm{d}s,$$

since $\eta|_L = 0$. Finally we obtain

$$\int_S\int \eta\left(\nabla^4 w - \frac{q}{D}\right)dS + \int_L \nabla^2 w \frac{\partial \eta}{\partial n} ds + \frac{1-\sigma}{2}\int_L \left\{\frac{\partial \eta}{\partial y}\frac{d}{ds}\left(\frac{\partial w}{\partial x}\right)\right.$$
$$\left. + \frac{\partial w}{\partial y}\frac{d}{ds}\left(\frac{\partial \eta}{\partial x}\right) - \frac{\partial \eta}{\partial x}\frac{d}{ds}\left(\frac{\partial w}{\partial y}\right) - \frac{\partial w}{\partial x}\frac{d}{ds}\left(\frac{\partial \eta}{\partial y}\right)\right\} ds = 0. \qquad (10)$$

It suffices to take the line integral in (10) over L_2, since the integral over L_1 vanishes as both functions η and w satisfy conditions (8). If the entire edge of the plate is rigidly fixed then the line integral in (10) vanishes. We suppose that $\eta(x, y)$ satisfies condition (8) not only on L_1 but also on the entire contour L. Then the line integral in (10) vanishes and this equation gives

$$\int_S\int \eta\left(\nabla^4 w - \frac{q}{D}\right)dS = 0.$$

By repeating the reasoning of § 11 we thus obtain the well-known differential equation for the equilibrium of a bent plate

$$\nabla^4 w = \frac{q}{D}, \qquad (11)$$

first obtained by Sophie Germain in 1815.

2. We now derive the second boundary condition for the freely resting part of the edge of the bent plate. In virtue of equation (11) the double integral in (10) vanishes, and, as has been stated, it suffices to take the line integral over L_2. Equation (10) assumes the form

$$\int_{L_2} \frac{\partial \eta}{\partial n}\nabla^2 w ds + \frac{1-\sigma}{2}\int_{L_2}\left\{\frac{\partial \eta}{\partial y}\frac{d}{ds}\left(\frac{\partial w}{\partial x}\right) + \frac{\partial w}{\partial y}\frac{d}{ds}\left(\frac{\partial \eta}{\partial x}\right)\right.$$
$$\left. - \frac{\partial \eta}{\partial x}\frac{d}{ds}\left(\frac{\partial w}{\partial y}\right) - \frac{\partial w}{\partial x}\frac{d}{ds}\left(\frac{\partial \eta}{\partial y}\right)\right\} ds = 0. \qquad (12)$$

At an arbitrarily chosen point on the curve L_2 we draw a tangent to it, which we direct so that the axes n and s are

orientated in the same sense as the axes x and y (Fig. 11) and we denote by ϑ the angle between the x and s axes. Further we put

$$\alpha = \cos\vartheta = \cos(s, x) = -\cos(n, y),$$
$$\beta = \sin\vartheta = \cos(s, y) = \cos(n, x).$$

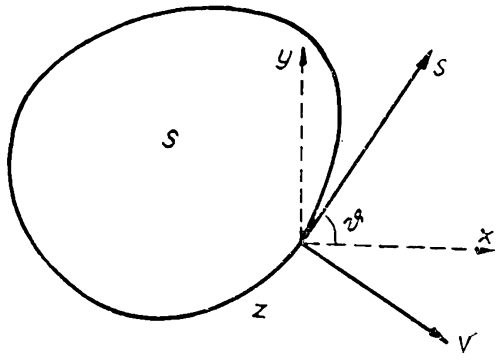

FIG. 11

The function w vanishes on the curve L_2, therefore on this curve also

$$\frac{\partial w}{\partial s} = \frac{\partial w}{\partial x}\cos(s, x) + \frac{\partial w}{\partial y}\cos(s, y) = 0.$$

On the other hand

$$\frac{\partial w}{\partial n} = \frac{\partial w}{\partial x}\cos(n, x) + \frac{\partial w}{\partial y}\cos(n, y).$$

Solving these two equations for $\frac{\partial w}{\partial x}$ and $\frac{\partial w}{\partial y}$, we find that on L_2

$$\frac{\partial w}{\partial x} = \beta\frac{\partial w}{\partial n}, \quad \frac{\partial w}{\partial y} = -\alpha\frac{\partial w}{\partial n}.$$

Similarly on L_2

$$\frac{\partial \eta}{\partial x} = \beta\frac{\partial \eta}{\partial n}, \quad \frac{\partial \eta}{\partial y} = -\alpha\frac{\partial \eta}{\partial n}.$$

Substituting this in (12) and carrying out the differentiation under the second integral sign, we obtain

$$\int_{L_2}\frac{\partial n}{\partial \eta}\left\{\nabla^2 w+(1-\sigma)\frac{\partial w}{\partial \eta}\left(\beta\frac{d\alpha}{ds}-\alpha\frac{d\beta}{ds}\right)\right\}ds=0.$$

But

$$\beta\frac{d\alpha}{ds}-\alpha\frac{d\beta}{ds}=-\frac{d\vartheta}{ds}=-\frac{1}{\varrho},$$

where ϱ is the radius of curvature of the curve L_2. Hence

$$\int_{L_2}\frac{\partial \eta}{\partial n}\left(\nabla^2 w-\frac{1-\sigma}{\varrho}\frac{\partial w}{\partial n}\right)ds=0.$$

Using the arbitrariness of $\frac{\partial \eta}{\partial n}$ on L_2 and applying the reasoning of § 11 to the given case we find the boundary condition in which we are interested in the form

$$\nabla^2 w-\frac{1-\sigma}{\varrho}\frac{\partial w}{\partial n}=0 \text{ on } L_2. \tag{13}$$

It is possible to show (although in fact we shall not do so) that condition (13) is the same as the condition for the absence of bending moments at the freely resting edge.

In analogous fashion, starting from the principle of minimum potential energy, it is possible to obtain the differential equations for the equilibrium of an elastic envelope and the corresponding boundary conditions (see the author's work [17]). The calculation in this case is naturally more complex. The principle of minimum potential energy can also be taken as the basis of a series of other approximate theories of the elastic state.

Without dwelling upon their derivation (see M. Sh. Birman [6]) we cite the boundary conditions for the free edge of the plate. We introduce the two-dimensional vectors

$$U=\left\{w|_L;-\left.\frac{\partial w}{\partial n}\right|_L\right\},\quad U_n=\left\{\left.\frac{\partial \nabla^2 w}{\partial n}\right|_L;\ \nabla^2 w|_L\right\}.$$

In terms of the vectors U and U_n the condition at the free edge of the plate has the form

$$U_n-(1-\sigma)\begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}s}\frac{1}{\varrho}\frac{\mathrm{d}}{\mathrm{d}s}, & \frac{\mathrm{d}^2}{\mathrm{d}s^2} \\ \frac{\mathrm{d}^2}{\mathrm{d}s^2}, & -\frac{1}{\varrho} \end{pmatrix} U=0.$$

3. It can be proved that the biharmonic operator entering into Sophie Germain's equation is symmetric and positive-bounded-below for the set of functions which satisfy the conditions (8) of rigid fixing on L_1, and the conditions (9) and (13) of free support on L_2. Here we confine ourselves to the two cases where the contour L of the plate is entirely rigidly fixed everywhere or freely supported everywhere.

Let the contour L be rigidly fixed. In order to verify the symmetry of the biharmonic operator in this case we suppose that the functions $w_1(x, y)$ and $w_2(x, y)$ both satisfy conditions (8). By formula (10) § 2

$$(\nabla^4 w_1, w_2)=\int_S\int w_2\nabla^4 w_1\mathrm{d}S$$

$$=\int_S\int\nabla^2 w_2\nabla^2 w_1\mathrm{d}S+\int_L\left(w_2\frac{\partial(\nabla^2 w_1)}{\partial n}-\nabla^2 w_1\frac{\partial w_2}{\partial n}\right)\mathrm{d}s.$$

The line integral vanishes in virtue of condition (8), thus

$$(\nabla^4 w_1, w_2)=\int_S\int\nabla^2 w_2\nabla^2 w_1\mathrm{d}S. \tag{14}$$

Since the right-hand side of formula (11) is symmetric in w_1 and w_2 $(\nabla^4 w_1, w_2)=(w_1, \nabla^4 w_2)$, i.e. the operator ∇^4 is symmetric for the set of functions satisfying conditions (8). We put $w_1=w_2=w$ in (14).

This gives

$$(\nabla^4 w, w)=\int_S\int(\nabla^2 w)^2\mathrm{d}S. \tag{15}$$

Now multiplying equation (6_2) by 2 and adding to (15), we obtain

$$(\nabla^4 w,\ w)=\int_S\int\left[\left(\frac{\partial^2 w}{\partial x^2}\right)^2+2\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2+\left(\frac{\partial^2 w}{\partial y^2}\right)^2\right]\mathrm{d}S. \qquad (16)$$

As we have already noted, conditions (8) are equivalent to

$$w=0,\ \frac{\partial w}{\partial x}=\frac{\partial w}{\partial y}=0 \quad \text{on} \quad L.$$

In such a case it is possible to apply Friedrichs's inequality (see § 22) to both first derivatives:

$$\int_S\int\left(\frac{\partial w}{\partial x}\right)^2\mathrm{d}S\leq\frac{1}{\varkappa}\int_S\int\left\{\left(\frac{\partial^2 w}{\partial x^2}\right)^2+\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2\right\}\mathrm{d}S,$$

$$\int_S\int\left(\frac{\partial w}{\partial y}\right)^2\mathrm{d}S\leq\frac{1}{\varkappa}\int_S\int\left\{\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2+\left(\frac{\partial^2 w}{\partial y^2}\right)^2\right\}\mathrm{d}S. \qquad (17)$$

Friedrichs's inequality can also be applied to the function w itself, since $w|_L=0$. Hence

$$\int_S\int w^2\mathrm{d}S\leq\frac{1}{\varkappa}\int_S\int\left\{\left(\frac{\partial w}{\partial x}\right)^2+\left(\frac{\partial w}{\partial y}\right)^2\right\}\mathrm{d}S$$

$$\leq\frac{1}{\varkappa^2}\int_S\int\left\{\left(\frac{\partial^2 w}{\partial x^2}\right)^2+2\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2+\left(\frac{\partial^2 w}{\partial y^2}\right)^2\right\}\mathrm{d}S,$$

or

$$(\nabla^4 w,\ w)\geq\varkappa^2\|w\|^2,$$

which also implies that the operator ∇^4 is positive-bounded-below. Hence it follows that the equilibrium problem of a rigidly fixed plate has a unique solution which can be constructed by using the methods of Chapter III. In particular if w_n, $n=1, 2, \ldots$ is some minimizing sequence constructed e.g. by the Ritz method it converges in energy to the exact solution. We now consider in greater detail the nature of the energy convergence in the given case. From formula (16) it follows that

$$\|w\|^2=\int_S\int\left\{\left(\frac{\partial^2 w}{\partial x^2}\right)^2+2\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2+\left(\frac{\partial^2 w}{\partial y^2}\right)^2\right\}\mathrm{d}S. \qquad (18)$$

Since the sequence of functions w_n converges in energy

$$\|w_n - w_m\| \underset{n,\, m\to\infty}{\longrightarrow} 0$$

or, *in full*

$$\int_S\int \left\{\left(\frac{\partial^2 w_n}{\partial x^2} - \frac{\partial^2 w_m}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 w_n}{\partial x \partial y} - \frac{\partial^2 w_m}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 w_n}{\partial y^2} - \frac{\partial^2 w_m}{\partial y^2}\right)^2\right\} dS \underset{n,\, m\to\infty}{\longrightarrow} 0.$$

Hence it clearly follows that

$$\left.\begin{aligned} \int_S\int \left(\frac{\partial^2 w_n}{\partial x^2} - \frac{\partial^2 w_m}{\partial x^2}\right)^2 dS &\underset{n,\, m\to\infty}{\longrightarrow} 0, \\ \int_S\int \left(\frac{\partial^2 w_n}{\partial x \partial y} - \frac{\partial^2 w_m}{\partial x \partial y}\right)^2 dS &\underset{n,\, m\to\infty}{\longrightarrow} 0, \\ \int_S\int \left(\frac{\partial^2 w_n}{\partial y^2} - \frac{\partial^2 w_m}{\partial y^2}\right)^2 dS &\underset{n,\, m\to\infty}{\longrightarrow} 0, \end{aligned}\right\} \tag{19}$$

i.e. the sequences of second derivatives of the approximate solutions converge in the mean.

The inequalities (4) and (5) now show that the sequences of first derivatives of w_n, and also of the w_n themselves, converge in the mean. We now prove that *the sequence* $\{w_n\}$ *converges uniformly within the closed domain* $\overline{\Omega}$, this being a special case of Sobolev's embedding theorem (see S. L. Sobolev [2]).

Conditions (8) are not natural but principal, therefore the functions w_n necessarily satisfy these conditions. We shall also assume that functions which are continuous with continuous first and second derivatives in $\overline{S} = S + L$ can always be secured. We enclose S within some rectangle Π (Fig. 8) and define the functions w in Π by taking them equal to zero outside S. Then these functions, and their first derivatives are

continuous in Π. We take an arbitrary point (x, y) in $\bar{S}$ and evaluate the integral

$$\int_0^x \int_0^y \frac{\partial^2 u(\xi, \eta)}{\partial \xi \partial \eta} \, d\xi d\eta, \quad u = w_n - w_m.$$

Performing the integration we find

$$\int_0^x \int_0^y \frac{\partial^2 u(\xi, \eta)}{\partial \xi \partial \eta} \, d\xi d\eta = u(x, y) - u(x, 0) - u(0, y) + u(0, 0).$$

The points $(x, 0)$, $(0, y)$ and $(0, 0)$ lie on the sides of the rectangle Π where $u = 0$, thus $u(x, 0) = u(0, y) = u(0, 0) = 0$ and consequently

$$u(x, y) = \int_0^x \int_0^y \frac{\partial^2 u}{\partial \xi \partial \eta} \, d\xi d\eta.$$

Hence

$$|u(x, y)| = \left| \int_0^x \int_0^y \frac{\partial^2 u}{\partial \xi \partial \eta} \, d\xi d\eta \right| \leq \int_{\Pi} \left| \frac{\partial^2 u}{\partial \xi \partial \eta} \right| d\xi d\eta,$$

or, since $u \equiv 0$ outside S.

$$|u(x, y)| \leq \int_S \left| \frac{\partial^2 u}{\partial \xi \partial \eta} \right| d\xi d\eta.$$

Now, by Cauchy's inequality,

$$u^2(x, y) \leq |S| \int_S \left(\frac{\partial^2 u}{\partial \xi \partial \eta} \right)^2 dS,$$

where $|S|$ denotes the area of the domain S. But $u = w_n - w_m$ therefore

$$|w_n(x, y) - w_m(x, y)| \leq \gamma(|S|) \left\{ \int_S \left(\frac{\partial^2 w_n}{\partial \xi \partial \eta} - \frac{\partial^2 w_m}{\partial \xi \partial \eta} \right)^2 dS \right\}^{\frac{1}{2}}.$$

By virtue of the inequality (19) the expression on the right tends to zero as $m, n \to \infty$.

By definition (cf. 1, **144**) the sequence w_n, $n=1, 2, \ldots$ converges uniformly in $\overline{S}$.

4. We now suppose that the contour L is freely supported so that the deflexion of the plate satisfies conditions (9) and (13). We consider the operator ∇^4, specified for the set of functions subjected to the conditions mentioned, and we shall prove that this operator is symmetric and positive-bounded-below.

Let $w_1(x, y)$ and $w_2(x, y)$ satisfy conditions (9) and (13) so that on L

$$w_1=w_2=0,\ \nabla^2 w_1-\frac{1-\sigma}{\varrho}\frac{\partial w_1}{\partial n}=\nabla^2 w_2-\frac{1-\sigma}{\varrho}\frac{\partial w_2}{\partial n}=0.$$

Then as in subsection **3** of this section,

$$(\nabla^4 w_1, w_2)=\int_S\int w_2\nabla^4 w_1\,\mathrm{d}S=\int_S\int \nabla^2 w_2\nabla^2 w_1\,\mathrm{d}S$$
$$+\int_L\left(w_2\frac{\partial(\nabla^2 w_1)}{\partial n}-\nabla^2 w_1\frac{\partial w_2}{\partial n}\right)\mathrm{d}s,$$

which by virtue of the boundary conditions assumes the form

$$(\nabla^4 w_1, w_2)=\int_S\int \nabla^2 w_2\nabla^2 w_1\,\mathrm{d}S-(1-\sigma)\int_L\frac{1}{\varrho}\frac{\partial w_1}{\partial n}\frac{\partial w_2}{\partial n}\mathrm{d}s. \quad (20)$$

We obtain the same result starting from the scalar product $(\nabla^4 w_2, w_1)$. Hence $(\nabla^4 w_1, w_2)=(\nabla^4 w_2, w_1)$, i. e. the operator considered is symmetric. In order to establish that it is positive-bounded-below we proceed in the following fashion.

(a) We carry out the steps of section 2 in the opposite order, substituting therein η by w_1 and w by w_2. We then find that

$$\int_L\frac{1}{\varrho}\frac{\partial w_1}{\partial n}\frac{\partial w_2}{\partial n}\,\mathrm{d}s=-\frac{1}{2}\int_L\left\{\frac{\partial w_1}{\partial y}\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\partial w_2}{\partial x}\right)\right.$$
$$\left.+\frac{\partial w_2}{\partial y}\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\partial w_1}{\partial y}\right)-\frac{\partial w_1}{\partial x}\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\partial w_2}{\partial y}\right)-\frac{\partial w_2}{\partial x}\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\partial w_1}{\partial y}\right)\right\}\mathrm{d}s.$$

We substitute this in (20) and put $w_1=w_2=w$.
This leads us to the equation

$$(\nabla^4 w,\ w)=\int_S\int(\nabla^2 w)^2\,\mathrm{d}S+(1-\sigma)\int_L\left\{\frac{\partial w}{\partial y}\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\partial w}{\partial x}\right)-\frac{\partial w}{\partial x}\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\partial w}{\partial y}\right)\right\}\mathrm{d}s,$$

or, making use of formula (6_1),

$$(\nabla^4 w,\ w)=\int_S\int\left\{\left(\frac{\partial^2 w}{\partial x^2}\right)^2+2\sigma\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}+\left(\frac{\partial^2 w}{\partial y^2}\right)^2+2(1-\sigma)\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2\right\}\mathrm{d}S. \tag{21}$$

It is interesting to note that the integral in (21) only differs by a constant factor $D/2$ from the expression [formula (5)]

$$\int\int_\Omega\int W(\mathbf{u})\mathrm{d}\Omega,$$

the potential energy of deformation of the disc.

(b) We have the identity

$$\xi^2+2\sigma\xi\eta+\eta^2-(1-\sigma)(\xi^2+\eta^2)=\sigma(\xi+\eta)^2.$$

By considering that $\sigma\geq 0$ we find that

$$\xi^2+2\sigma\xi\eta=\eta^2\geq(1-\sigma)(\xi^2+\eta^2).$$

Putting $\xi=\dfrac{\partial^2 w}{\partial x^2}$, $\eta=\dfrac{\partial^2 w}{\partial y^2}$ we obtain from (21) the inequality

$$(\nabla^4 w,\ w)\geq(1-\sigma)\int_S\int\left\{\left(\frac{\partial^2 w}{\partial x^2}\right)^2+2\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2+\left(\frac{\partial^2 w}{\partial y^2}\right)\right\}\mathrm{d}S. \tag{22}$$

We obtain an analogous inequality by the substitution of the factor $1-\sigma$ for $1+\sigma$ if $\sigma<0$.

c) We apply to the derivative $\frac{\partial w}{\partial x}$ the Poincaré inequality [formula (24) § 22], which in this case gives

$$\int_S\int\left(\frac{\partial w}{\partial x}\right)^2 dS \leq A\int_S\int\left\{\left(\frac{\partial^2 w}{\partial x^2}\right)^2+\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2\right\}dS$$

$$+B\left(\int_S\int\frac{\partial w}{\partial x}\,dS\right)^2. \tag{23}$$

By Gauss's theorem

$$\int_S\int\frac{\partial w}{\partial x}\,dS=\int_L w\cos(n,\,x)ds=0, \tag{24}$$

and the inequality (23) assumes the form:

$$\int_S\int\left(\frac{\partial w}{\partial x}\right)^2 dS \leq A\int_S\int\left\{\left(\frac{\partial^2 w}{\partial y^2}\right)^2+\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2\right\}dS.$$

Similarly, we find

$$\int_S\int\left(\frac{\partial w}{\partial y}\right)^2 dS \leq A\int_S\int\left\{\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2+\left(\frac{\partial^2 w}{\partial y^2}\right)^2\right\}dS.$$

Adding the latter inequalities we obtain

$$\int_S\int\left\{\left(\frac{\partial w}{\partial x}\right)^2+\left(\frac{\partial w}{\partial y}\right)^2\right\}dS \leq A\int_S\int\left\{\left(\frac{\partial^2 w}{\partial x^2}\right)^2+2\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2\right.$$

$$\left.+\left(\frac{\partial^2 w}{\partial y^2}\right)^2\right\}dS.$$

Comparison with the inequality (22) gives

$$(\nabla^4 w,\ w)\geq\frac{1-\sigma}{A}\int_S\int\left\{\left(\frac{\partial w}{\partial x}\right)^2+\left(\frac{\partial w}{\partial y}\right)^2\right\}dS. \tag{25}$$

Condition (9) allows us to apply Friedrichs's inequality to the function w:

$$\int_S \int \left\{ \left(\frac{\partial w}{\partial x} \right)^2 + \left(\frac{\partial w}{\partial y} \right)^2 \right\} dS \geq \varkappa \int_S \int w^2 dS = \varkappa \|w\|^2.$$

Substituting this in (25), we arrive at the inequality

$$(\nabla^4 w, w) \geq \gamma^2 \|w\|^2, \quad \gamma^2 = \frac{(1-\sigma)\varkappa}{A}, \tag{26}$$

which also means that the operator considered is positive-bounded-below.

We leave for the reader the formulation of the consequences stemming from this fact. We merely note that in this case it is possible to prove the uniform convergence of the minimizing sequence if S. L. Sobolev's "embedding theorem", mentioned above, is used. However, such a proof lies beyond the scope of the present book.

5. The plane problem of elasticity theory, for an elastic medium under specified external forces in the case when the domain S is finite and simply-connected, reduces to a problem similar to the variational problem for the case of a plate rigidly fixed at its edge. In this case, as is known, it is possible to reduce the problem to the integration of the biharmonic equation

$$\nabla^4 u = 0 \tag{27}$$

under boundary conditions

$$u\Big|_L = f_1(s), \quad \frac{\partial u}{\partial n}\Big|_L = f_2(s) \tag{28}$$

where $f_1(s)$ and $f_2(s)$ are known functions. If $\psi(x, y)$ is an arbitrary four times continuously differentiable function satisfying conditions (28), then putting $u - \psi = w$, we find that w satisfies boundary conditions (8) and equation (11) in which $q/D = -\nabla^4 \psi$. The function w can be found by solving the problem of the minimum of the functional

$$F(w) = (\nabla^4 w, w) + 2(\nabla^4 \psi, w).$$

Putting $w=u-\psi$ here, we readily find

$$F(w)=\Phi(u)+\Phi(\psi)-2\int_S\int\left(\frac{\partial^2 u}{\partial x^2}\frac{\partial^2\psi}{\partial x^2}+2\frac{\partial^2 u}{\partial x\partial y}\frac{\partial^2\psi}{\partial x\partial y}\right.$$
$$\left.+\frac{\partial^2 u}{\partial y^2}\frac{\partial^2\psi}{\partial y^2}\right)\mathrm{d}S+2\int_S\int u\nabla^4\psi\,\mathrm{d}S-2\int_S\int\psi\nabla^2\psi\,\mathrm{d}S, \qquad (29)$$

where for brevity we have taken

$$\Phi(u)=\int_S\int\left\{\left(\frac{\partial^2 u}{\partial x^2}\right)^2+2\left(\frac{\partial^2 u}{\partial x\partial y}\right)^2+\left(\frac{\partial^2 u}{\partial y^2}\right)^2\right\}\mathrm{d}S.$$

We note that on the right-hand side of equation (29) the second and fifth terms are constant. We integrate the integral in the third term twice by parts. From formulae (28) it follows that $\left.\frac{\partial u}{\partial x}\right|_L$ and $\left.\frac{\partial u}{\partial y}\right|_L$ are known, therefore the line integrals arising from the integration by parts are known constants. With this in mind, we find

$$\int_S\int\left(\frac{\partial^2 u}{\partial x^2}\frac{\partial^2\psi}{\partial x^2}+2\frac{\partial^2 u}{\partial x\partial y}\frac{\partial^2\psi}{\partial x\partial y}+\frac{\partial^2 u}{\partial y^2}\frac{\partial^2\psi}{\partial y^2}\right)\mathrm{d}S$$
$$=\int_S\int u\nabla^4\psi\,\mathrm{d}S+\text{const.}$$

Substituting this in (29) we find that

$$F(w)=\Phi(u)+\text{const.}$$

Thus, the plane problem of elasticity theory under the conditions mentioned above reduces to the problem of finding the minimum of the functional $\Phi(u)$ for the set of functions $u(x, y)$ which together with their normal derivatives assume the specified values (28).

In finding the minimum it is possible to replace $\Phi(u)$ by the simpler functional

$$\Phi_0(u)=\int_S\int(\nabla^2 u)^2\mathrm{d}S.$$

In fact, by formula (6)

$$\Phi_0(u)-\Phi(u)=2\int_L \frac{\partial u}{\partial x}\,\mathrm{d}\left(\frac{\partial u}{\partial y}\right),$$

which, in virtue of boundary conditions (28) is a known constant.

§ 28. The bending of thin plates to which both normal loads and loads acting in the central plane are applied

Let the plate be under the simultaneous influence of a normal load $q(x, y)$ and the stresses T_x, T_{xy}, T_y, which act in the central plane of the disc. We shall consider that volume forces are absent. Then T_x, T_{xy}, T_y are related by the homogeneous equilibrium equations,

$$\frac{\partial T_x}{\partial x}+\frac{\partial T_{xy}}{\partial y}=0; \qquad \frac{\partial T_{xy}}{\partial x}+\frac{\partial T_y}{\partial y}=0. \tag{*}$$

If the deflexion $w(x, y)$ of the plate is small, it can be considered that it approximately satisfies the well-known equation (see e.g. P. F. Papkovich [1], p. 252 or S. P. Timoshenko [2])

$$\nabla^4 w-\frac{h}{D}\left[T_x\frac{\partial^2 w}{\partial x^2}+2T_{xy}\frac{\partial^2 w}{\partial x\partial y}+T_y\frac{\partial^2 w}{\partial y^2}\right]=\frac{q}{D}. \tag{1}$$

As in the previous section, h denotes the thickness of the plate and D is the rigidity. Just as in § 27 the boundary conditions for the rigidly fixed part of the edge have the form

$$w=0, \qquad \frac{\partial w}{\partial n}=0, \tag{2}$$

and at the freely supported part of the edge

$$w=0, \quad \nabla^2 w-\frac{1-\sigma}{\varrho}\frac{\partial w}{\partial n}=0. \tag{3}$$

We shall prove that the operator on the left-hand side of equation (1) is symmetric for each of the sets of functions satisfying either conditions (2) or conditions (3). (M.Sh. Birman drew the author's attention to this fact.) The symmetry of the biharmonic operator was established in § 27. In order to prove our assertion it suffices to satisfy ourselves that the operator

$$Tw = T_x \frac{\partial^2 w}{\partial x^2} + 2T_{xy} \frac{\partial^2 w}{\partial x \partial y} + T_y \frac{\partial^2 w}{\partial y^2} \tag{4}$$

is symmetric with respect to the set of functions satisfying the boundary condition

$$w|_L = 0, \tag{5}$$

which is satisfied in both the cases indicated above. To this end we note that, using the equilibrium equation (*), it is possible to transform the operator T to the form

$$Tw = \frac{\partial}{\partial x}\left(T_x \frac{\partial w}{\partial x} + T_{xy} \frac{\partial w}{\partial y}\right) + \frac{\partial}{\partial y}\left(T_{xy} \frac{\partial w}{\partial x} + T_y \frac{\partial w}{\partial y}\right).$$

We now construct the scalar product (Tw_1, w_2), assuming for this that both the functions w_1, w_2 satisfy boundary conditions (5). Integrating by parts and bearing in mind that, by virtue of the condition mentioned, the line integral vanishes, we find

$$(Tw_1, w_2) = \int_S \int w_2 \left[\frac{\partial}{\partial x}\left(T_x \frac{\partial w_1}{\partial x} + T_{xy} \frac{\partial w_1}{\partial y}\right)\right.$$

$$\left. + \frac{\partial}{\partial y}\left(T_{xy} \frac{\partial w_1}{\partial x} + T_y \frac{\partial w_1}{\partial y}\right)\right] dS = -\int_S \int \left[T_x \frac{\partial w_1}{\partial x} \frac{\partial w_2}{\partial x}\right.$$

$$\left. + T_{xy}\left(\frac{\partial w_1}{\partial x} \frac{\partial w_2}{\partial y} + \frac{\partial w_1}{\partial y} \frac{\partial w_2}{\partial x}\right) + T_y \frac{\partial w_1}{\partial y} \frac{\partial w_2}{\partial y}\right] dS. \tag{6}$$

Since w_1 and w_2 enter symmetrically into the latter integral the same result is obtained obviously, if we start from the scalar product (w_1, Tw_2). Hence $(Tw_1, w_2) = (w_1, Tw_2)$, which was to be proved.

THEOREM 1. *If the stresses* T_x, T_{xy}, T_y, *are sufficiently small in absolute value, the operator on the left-hand side of equation* (1) *is positive-bounded-below for each of the sets of functions which satisfy over the entire contour either conditions* (2) *or conditions* (3).

We construct the scalar product

$$\left(\nabla^4 w - \frac{h}{D} Tw,\ w\right) = (\nabla^4 w,\ w) - \frac{h}{D}(Tw,\ w). \tag{7}$$

If the entire edge of the plate is rigidly fixed [conditions (2)], then by formulae (16) and (18) § 27

$$(\nabla^4 w,\ w) \geq \varkappa \int_S \int \left\{ \left(\frac{\partial w}{\partial x}\right)^2 + \left(\frac{\partial w}{\partial y}\right)^2 \right\} \mathrm{d}S.$$

If the entire edge is freely supported [conditions (3)] then by formula (25) § 27

$$(\nabla^4 w,\ w) \geq \frac{1-\sigma}{A} \int_S \int \left\{ \left(\frac{\partial w}{\partial x}\right)^2 + \left(\frac{\partial w}{\partial y}\right)^2 \right\} \mathrm{d}S.$$

Denoting by C the quantity $\varkappa$ in the case of a rigidly fixed edge and the quantity $\frac{1-\sigma}{A}$ in the case of a freely supported edge, we have in both cases

$$(\nabla^4 w,\ w) \geq C \int_S \int \left\{ \left(\frac{\partial w}{\partial x}\right)^2 + \left(\frac{\partial w}{\partial y}\right)^2 \right\} \mathrm{d}S. \tag{8}$$

We now estimate the quantity $(Tw,\ w)$. In formula (6) we put $w_1 = w_2 = w$.

We then obtain

$$(Tw,\ w) = -\int_S \int \left[T_x \left(\frac{\partial w}{\partial x}\right)^2 + 2T_{xy} \frac{\partial w}{\partial x} \frac{\partial w}{\partial y} + T_y \left(\frac{\partial w}{\partial y}\right)^2 \right] \mathrm{d}S. \tag{9}$$

Let N be the greatest of the maxima of the quantities T_x, T_{xy}, T_y so that

$$|T_x| \leq N, \quad |T_{xy}| \leq N, \quad |T_y| \leq N.$$

Then

$$|(Tw, w)| \leq N \int_S \int \left\{ \left| \frac{\partial w}{\partial x} \right| + \left| \frac{\partial w}{\partial y} \right| \right\}^2 dS,$$

or, if we use the inequality $(a+b)^2 \leq 2(a^2+b^2)$

$$|(Tw, w)| \leq 2N \int_S \int \left[\left(\frac{\partial w}{\partial x} \right)^2 + \left(\frac{\partial w}{\partial y} \right)^2 \right] dS. \tag{10}$$

Inequality (10) still holds if by $2N$ is understood the maximum of the largest of the principal normal tensions which act in the plane of the plate.

With the aid of relations (8) and (10) we obtain from formula (7)

$$\left(\nabla^4 w - \frac{h}{D} Tw, w \right) \geq \left(C - \frac{2Nh}{D} \right) \int_S \int \left\{ \left(\frac{\partial w}{\partial x} \right)^2 + \left(\frac{\partial w}{\partial y} \right)^2 \right\} dS. \tag{11}$$

Since $w|_L = 0$, Friedrichs's inequality is valid

$$\int_S \int \left\{ \left(\frac{\partial w}{\partial x} \right)^2 + \left(\frac{\partial w}{\partial y} \right)^2 \right\} dS \geq \varkappa \|w\|^2.$$

Let the stresses T_x, T_{xy}, T_y be so small that

$$N < \frac{CD}{2h}. \tag{12}$$

Then the right-hand side of inequality (11) is positive and we reinforce the inequality by replacing the integral on the right by $\varkappa \|w\|^2$. Then

$$\left(\nabla^4 w - \frac{h}{D} Tw, w \right) \geq \left(C - \frac{2Nh}{D} \right) \varkappa \|w\|^2, \tag{13}$$

and theorem 1 is valid if the maximum of the stresses satisfies inequality (12).

From the theorem just proved and also from the results of § 27 and Ch. III in particular §§ 11, 19, 12–15 the following conclusions hold.

Let the plate be under the simultaneous influence of some normal load $q(x, y)$ and of stresses T_x, T_{xy}, T_y acting in the central plane of the plate, and let the entire edge of the plate be either rigidly fixed or freely supported. If the maximum of the stresses T_x, T_{xy}, T_y satisfies inequality (12) *then the equilibrium problem for the plate has a solution which is unique and which can be obtained as the function minimizing the functional*

$$F(w)=\int_S\int\left\{\left(\frac{\partial^2 w}{\partial x^2}\right)^2+2\sigma\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}+\left(\frac{\partial^2 w}{\partial y^2}\right)^2+2(1-\sigma)\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2\right.$$

$$\left.+\frac{h}{D}\left[T_x\left(\frac{\partial w}{\partial x}\right)^2+2T_{xy}\frac{\partial w}{\partial x}\frac{\partial w}{\partial y}+T_y\left(\frac{\partial w}{\partial y}\right)^2\right]-\frac{2q}{D}w\right\}\mathrm{d}S. \quad (14)$$

As will be proved in § 38, this problem does not have a solution generally. If the edge of the plate is rigidly fixed, functional (14) can be simplified and reduced to either of the two forms:

$$F(w)=\int_S\int\left\{(\nabla^2 w)^2+\frac{h}{D}\left[T_x\left(\frac{\partial w}{\partial x}\right)^2+2T_{xy}\frac{\partial w}{\partial x}\frac{\partial w}{\partial y}\right.\right.$$

$$\left.\left.+T_y\left(\frac{\partial w}{\partial y}\right)^2\right]-\frac{2q}{D}w\right\}\mathrm{d}S; \quad (15_1)$$

$$F(w)=\int_S\int\left\{\left(\frac{\partial^2 w}{\partial x^2}\right)^2+2\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2+\left(\frac{\partial^2 w}{\partial y^2}\right)^2+\frac{h}{D}\left[T_x\left(\frac{\partial w}{\partial x}\right)^2\right.\right.$$

$$\left.\left.+2T_{xy}\frac{\partial w}{\partial x}\frac{\partial w}{\partial y}+T_y\left(\frac{\partial w}{\partial y}\right)^2\right]-\frac{2q}{D}w\right\}\mathrm{d}S. \quad (15_2)$$

The identity of the functionals (14) (15_1) (15_2) for a rigidly fixed edge follows from the formula (6_2) § 27.

The solution of the problem of the minimum of the functional $F(w)$ can be constructed as the limit of a minimizing sequence, in particular the sequence of Ritz's approximate solutions, and also in the form of orthogonal series (§ 12). If w_n, $n=1, 2, \ldots$ is a minimizing sequence and w is the exact solution, then $w_n \to w$ uniformly in $\bar{S}=S+L$ and the first and second

derivations of w_n converge in the mean to the corresponding derivative of w.

Making use of the "embedding theorem" frequently cited above, it is possible to prove that the operator $\nabla^4 - \frac{h}{D}T$ is positive-bounded-below under the condition of type (12), in the case where the edge of the plate is partly rigidly fixed and partly freely supported, and also under certain other fixing conditions.

We note one simple and important case where the operator $\nabla^4 - \frac{h}{D}T$ is positive-bounded-below and all the conclusions just enumerated occur although inequality (12) need not be satisfied. This is the case when at any point in the plate *the set of stresses* T_x, T_{xy}, T_y *reduce to tensions in both principal directions*. As before we assume that the edge of the disc is either entirely rigidly fixed or entirely freely supported. In order to prove our assertion it suffices to establish that $(-Tw, w) \geq 0$ as long as $w|_L = 0$.

First of all we note that the expression under the integral sign in (9) is invariant under a rotation of the coordinate axes; of this we can easily convince ourselves by means of simple, albeit somewhat tedious transformations. With this in mind, we take an arbitrary point in S and rotate the coordinate axes so that they coincide with the principal axes of the set of stresses T_x, T_{xy}, T_y at the point chosen. Then at this point the integrand in (9) assumes the form

$$T_1\left(\frac{\partial w}{\partial x_1}\right)^2 + T_2\left(\frac{\partial w}{\partial x_2}\right)^2,$$

where x_1, x_2, are the principal axes of the stresses and T_1, T_2, are the principal normal stresses. By assumption the stresses T_1, T_2, are tensions and, consequently, positive. Hence it follows that at the chosen point in S the integrand in (9) is non-negative. Since the point was chosen arbitrarily, the

function is also non-negative in the whole domain. Hence it follows that

$$(-Tw, w) \geq 0$$

and

$$(\nabla^4 w - \frac{h}{D} Tw, w) \geq (\nabla^4 w, w) \geq C\|w\|^2,$$

which was to be proved.

CHAPTER V

THE EIGENVALUE PROBLEM

§ 29. The eigenvalue problem; its connexion with characteristic vibrations and stability of a system

An important role in mathematical physics is played by the concepts of *eigenvalues and eigenfunctions* of an equation. Let the equation

$$Au - \lambda Bu = 0, \tag{1}$$

be given, where A and B are certain linear operators and λ is a numerical factor. In virtue of the properties of linear operators (see § 4) $AO = 0$ and $BO = 0$; hence it follows that the function $u \equiv 0$ satisfies equation (1), whatever the factor λ. The function $u \equiv 0$ is commonly called the *trivial solution* of equation (1). We now pose a problem: to find those values of the parameter λ for which equation (1) has a solution, different from identically zero. Such values of λ, if they exist, are called *the characteristic values* or *eigenvalues* of equation (1), and the non-trivial (i.e. different from identically zero) solutions of this equation corresponding to them are known as its eigenfunctions. Thus, if the number λ_0 and the function $u_0(P)$ are such that $u_0(P) \not\equiv 0$ and

$$Au_0 - \lambda_0 Bu_0 = 0,$$

then λ_0 is an eigenvalue of equation (1) and $u_0(P)$ is the eigenfunction which *corresponds to* the eigenvalue λ_0.

Especially important is the particular case when B is the identity operator: $Bu = u$. In this case equation (1) assumes the form

$$Au - \lambda u = 0. \tag{2}$$

The eigenfunctions and eigenvalues of equation (2) are also known as *eigenfunctions and eigenvalues of the operator A*. The set of all the eigenvalues of an operator is known as its spectrum.

Example. Let $Au=-\frac{d^2u}{dx^2}$, $0<x<l$, the functions from the field of definition of the operator A being subjected to the boundary conditions

$$u(0)=u(l)=0. \tag{3}$$

The eigenvalues of operator A are, in the given case, the values λ for which the equation

$$\frac{d^2u}{dx^2}+\lambda u=0 \tag{4}$$

has a solution which satisfies conditions (3) and is different from identically zero; such a solution is an eigenfunction of our operator. In the example given, it is easy to find the eigenfunctions and eigenvalues of the operator. The eigenfunctions satisfy differential equation (4), the general solution of which has the form

$$u(x)=C\sin\sqrt{(\lambda)}x+C_1\cos\sqrt{(\lambda)}x.$$

From the condition $u(0)=0$ it follows that $C_1=0$ and $u(x)=C\sin\sqrt{(\lambda)}x$ and the condition $u(l)=0$ gives $C\sin\sqrt{(\lambda)}l=0$. If $C=0$, $u\equiv 0$ and we do not obtain an eigenfunction. Therefore, it must follow that

$$\sin\sqrt{(\lambda)}l=0,$$

whence we readily find the eigenvalues

$$\lambda_n=\frac{\pi^2n^2}{l^2},\quad n=1,2,\ldots$$

and the eigenfunctions corresponding to them

$$u_n(x)=C_n\sin\frac{\pi nx}{l},\quad n=1,\ 2,\ldots\ .$$

The quantities C_n are different from zero and are otherwise arbitrary and can, in particular, be different for different n.

If $u(P)$ is an eigenfunction of equation (1) and C is any non-zero constant, then $Cu(P)$ is also an eigenfunction; in general if $u_1(P)$, $u_2(P)$, . . . , $u_k(P)$ are eigenfunctions of equation (1), then their linear combination

$$C_1u_1(P)+C_2u_2(P)+\ \ldots\ +C_ku_k(P),$$

is an eigenfunction of the same equation (providedthat this combination differs from identically zero) for any constants $C_1, C_2, \ldots, C_k$. It therefore, suffices to know only the linearly independent eigenfunctions corresponding to a given eigenvalue.

The number of linearly independent functions corresponding to a given eigenvalue is known as the rank of the eigenvalue. The rank of an eigenvalue can also be infinite. Problems of the characteristic vibrations of mechanical systems usually reduce to the seeking of eigenvalues and eigenfunctions.

We consider as an example a uniform stretched string of length l which in the equilibrium state occupies the segment $0 \leq x \leq l$ of the x-axis. The transverse vibrations of such a string are defined by the equation (cf. 2, **163**)

$$\frac{\partial^2 U}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 U}{\partial t^2}; \quad c = \sqrt{\left(\frac{T}{\varrho}\right)}. \tag{5}$$

Here $U(x, t)$ is the displacement of the string from the equilibrium position at the point x and at the instant of time t, T is the tension in the string and ϱ is its mass per unit length. It is assumed that the string is not acted upon by any external forces, so that the nature of vibrations is dependent only on the initial displacement and the initial momentum imparted to the string. If the ends of the string are firmly fixed, then

$$U(0, t) = U(l, t) = 0. \tag{6}$$

Vibrations for which the displacement $U(x, t)$ has the form

$$U(x, t) = u(x) \begin{matrix} \cos \\ \sin \end{matrix} \omega t, \quad \omega = \text{const.}$$

are known as the characteristic vibrations of the string. Substituting this in (5) and (6) we find that $u(x)$ satisfies the differential equation

$$\frac{\mathrm{d}^2 u}{\mathrm{d}x^2} + \lambda u = 0; \quad \lambda = \frac{\omega^2}{c^2} = \frac{\omega^2 \varrho}{T}$$

and the boundary conditions

$$u(0)=u(l)=0.$$

From the nature of the problem $u(x)\not\equiv 0$, otherwise there would be no vibration of the string. The problem of the characteristic vibrations of the string is thus reduced to the problem, which we have already solved, of the eigenvalues and eigenfunctions of the operator $-\frac{d^2}{dx^2}$ under the boundary conditions (3). The characteristic vibrations of the string are defined to within a constant factor by the formula:

$$U_n(x,\,t)=\sin\frac{\pi n x}{l}\cos\frac{c\pi n t}{l}\sin\frac{}{}.$$

If the string is non-uniform, then $\varrho=\varrho(x)$ is a function of x. In this case the problem of the characteristic vibrations of the string leads to the same boundary conditions (3) and to the equation

$$\frac{d^2u}{dx^2}+\lambda\varrho(x)u=0, \tag{7}$$

where this time $\lambda=\frac{\omega^2}{T}$. Equation (7) is a particular case of equation (1) in which $A=-\frac{d^2}{dx^2}$ and B is the operator of multiplication by the function $\varrho(x)$, both operators being defined for functions which vanish at the ends of the string.

We consider further the characteristic vibrations of a membrane. If the membrane is not subjected to external forces then the equation for its transverse vibrations has the form (cf. 2, **176**)

$$\nabla^2U=\frac{\partial^2U}{\partial x^2}+\frac{\partial^2U}{\partial y^2}=\frac{1}{c^2}\frac{\partial^2U}{\partial t^2}. \tag{8}$$

Let the membrane in the equilibrium state occupy the domain Ω in the xy-plane, which is bounded by the contour S. If the edge of the membrane is firmly fixed

$$U|_S=0. \tag{9}$$

The characteristic vibrations of the membrane have the form

$$U(x, y, t) = u(x, y) \begin{matrix} \cos \\ \sin \end{matrix} \omega t;$$

and finding them reduces to the integration of the equation

$$\nabla^2 u + \lambda u = 0, \quad \lambda = \frac{\omega^2}{c^2} \tag{10}$$

in the domain Ω, under the boundary condition

$$u|_S = 0. \tag{11}$$

Since necessarily $u(x, y) \not\equiv 0$ our problem reduces to finding the eigenvalues and eigenfunctions of the operator $-\nabla^2$ under the boundary condition (11). This problem is considerably more difficult than the analogous problem for a string and its solution becomes comparatively simple in a few cases, such as e.g. the case of a rectangular or circular membrane (cf. 2, **177** and **178**).

A further class of problems which also reduce to the problem of eigenvalues and eigenfunctions is also well known. This is the problem of the stability of mechanical systems. We shall not formulate this problem in general form but we shall elucidate it by an example. We consider the bending of a thin plate upon which there act forces which are applied to its central plane and which are proportional to some parameter λ; let these forces coincide with the stresses λT_x, λT_{xy}, λT_y. Let the normal load be absent and for precision, let the edge of the plate be rigidly fixed. The deflexion of the plate is defined (see § 28) by the homogeneous (since $q(x, y) \equiv 0$) differential equation

$$\nabla^4 w - \frac{\lambda h}{D}\left[T_x \frac{\partial^2 w}{\partial x^2} + 2T_{xy} \frac{\partial^2 w}{\partial x \partial y} + T_y \frac{\partial^2 w}{\partial y^2}\right] = 0 \tag{12}$$

and by the homogeneous boundary conditions

$$w|_L = 0, \quad \left.\frac{\partial w}{\partial n}\right|_L = 0. \tag{13}$$

The system of equations (12) and (13) has the trivial solution $w=0$ which obviously means that an unbent equilibrium shape of the plate is possible. If λ is sufficiently small, (see formula (12) § 28), i.e. if

$$|\lambda| < \frac{CD}{2Nh} \tag{14}$$

where N is the greatest of the maxima of the stresses T_x, T_{xy}, T_y, then the trivial solution is unique, which means that the unbent equilibrium shape is unique. It is clear in this case that a sufficiently small change in λ does not violate inequality (14) and consequently does not violate the unbent plane equilibrium shape which thus remains *stable* with respect to small changes of the parameter λ. If λ does not satisfy inequality (14) then it is not excluded that at a certain value of λ (known as the *critical* value) equation (12) may have a *non-trivial* solution which satisfies boundary conditions (13). In this case, besides the planar, a bent equilibrium shape is also possible. At a value of λ close to critical the plane equilibrium state becomes unstable. In this connexion the seeking of the critical values of the parameter (especially of the lowest critical value) becomes important; this, in turn, reduces to seeking the eigenvalues of equation (12) under boundary conditions (13).

In the general case the stability problem reduces to the problem of the eigenvalues of a linear equation, starting from the non-linear equations of the given problem. The rigorous mathematical foundation of such a procedure is given in the works of M. A. Krasnosel'sky [2], [3]. For elastic systems such a stability problem is discussed at length in the book by V. V. Novozhilov [1] which, starting from the non-linear equations of elasticity theory, arrives at the problem of the eigenvalues of certain linear equations (or of a system of equations).

§ 30. Eigenvalues and eigenfunctions of a symmetric operator

In practical questions the investigation of the eigenvalues and eigenfunctions of a *symmetric* operator is especially important; some of their simple and important properties will be established in the present section.

THEOREM 1. *The eigenvalues of a symmetric operator are real numbers.*

Let λ_0 and $\varphi_0(P)$ be an eigenvalue and its corresponding eigenfunction of the operator A.

Then we have the identity

$$A\varphi_0 = \lambda_0\varphi_0.$$

If the function φ_0 is real, then multiplying this identity scalarly by φ_0 and then dividing by the positive quantity $(\varphi_0, \varphi_0) = \|\varphi_0\|^2$ we obtain

$$\lambda_0 = \frac{(A\varphi_0, \varphi_0)}{\|\varphi_0\|^2}, \tag{1}$$

and the real nature of the number λ_0 is obvious.

Now let $\varphi_0(P)$ be complex, $\varphi_0(P) = \psi(P) + i\omega(P)$ where $\psi(P)$ and $\omega(P)$ are real. Then

$$A(\psi + i\omega) = \lambda_0(\psi(P) + i\omega(P)), \tag{2}$$

or

$$A\psi + iA\omega = \lambda_0(\psi(P) + i\omega(P)).$$

We multiply (in the general meaning of this word) the latter equality by $\psi(P) - i\omega(P)$ and then integrate over the given domain Ω. This leads us to the equation

$$\int_\Omega [\psi A\psi + \omega A\omega + i(\psi A\omega - \omega A\psi)]\,d\Omega = \lambda_0 \int_\Omega [\psi^2(P) + \omega^2(P)]\,d\Omega,$$

which can also be presented in the form:

$$(A\psi, \psi) + (A\omega, \omega) + i[(A\omega, \psi) - (\omega, A\psi)] = \lambda_0(\|\psi\|^2 + \|\omega\|^2).$$

Since the operator A is symmetric

$$(A\omega, \psi) = (\omega, A\psi),$$

and the left-hand side of the latter equality is real.
Further $\varphi_0(P) \not\equiv 0$. Hence $\psi^2(P) + \omega^2(P) = |\varphi_0(P)|^2 \not\equiv 0$ and consequently

$$\|\psi\|^2 + \|\omega\|^2 = \int_{\Omega} |\varphi_0(P)|^2 d\Omega > 0.$$

Now

$$\lambda_0 = \frac{(A\psi, \psi) + (A\omega, \omega)}{\|\psi\|^2 + \|\omega\|^2}$$

and it is clear that λ_0 is a real number.

We note that the eigenfunctions of a symmetric operator can be assumed to be real.

In fact, if $\varphi_0(P) = \psi(P) + i\omega(P)$ then, separating the real and imaginary parts in (2) we obtain,

$$A\psi = \lambda_0 \psi, \qquad A\omega = \lambda_0 \omega,$$

and the real eigenfunctions ψ and ω correspond to the eigenvalue λ_0. From now on we shall always consider that the eigenfunctions of a symmetric operator are real.

NOTE: Once the eigenfunctions of a symmetric operator are assumed real, formula (1) is always true for the eigenvalues of such an operator.

Hence, incidentally, it follows immediately that all the eigenvalues of a positive definite and, consequently, of a positive-bounded-below operator are positive.

THEOREM 2. *The eigenfunctions of a symmetric operator which correspond to different eigenvalues are mutually orthogonal.*

Let λ_1 and λ_2 be two unequal eigenvalues of the symmetric operator A, and $\varphi_1(P)$, $\varphi_2(P)$ be the eigenfunctions corresponding to them. Then

$$A\varphi_1 = \lambda_1 \varphi_1, \qquad A\varphi_2 = \lambda_2 \varphi_2.$$

We multiply the first equation scalarly by φ_2 and the second by φ_1.

Subtracting we obtain

$$(A\varphi_1, \varphi_2)-(\varphi_1, A\varphi_2)=(\lambda_1-\lambda_2)(\varphi_1, \varphi_2).$$

The left-hand side of the latter equation equals zero, because the operator A is symmetric; also $\lambda_1-\lambda_2\neq 0$ and so $(\varphi_1, \varphi_2)=0$.

If several eigenfunctions correspond to a given eigenvalue, they can be made orthogonal by using an orthogonalization process (§ 10).

This permits us to assume without reservation from now on that *the set comprising of all the eigenfunctions of a symmetric operator constitutes an orthonormal system.*

THEOREM 3. *The system of eigenfunctions of a positive operator is orthogonal in energy.*

Let λ_1 and φ_1 be an eigenvalue and an eigenfunction of the positive definite operator A such that

$$A\varphi_1=\lambda_1\varphi_1 \tag{3}$$

and let φ_2 be an eigenfunction of the same operator, but different from φ_1. As stated above, φ_1 and φ_2 are orthogonal in the ordinary sense: $(\varphi_1, \varphi_2)=0$. Multiplying equation (3) scalarly by φ_2, we find $(A\varphi_1, \varphi_2)=0$, which was to be proved.

N O T E: It often happens that the system of eigenfunctions is not only orthogonal but also complete in energy. Then this system can be used to construct the solution of the equation

$$Au=f \tag{4}$$

in the form of an orthogonal series, as in § 12.

Let the system of eigenfunctions of a positive-bounded-below operator A be complete in energy and, in correspondence with what was stated previously, orthonormal in the ordinary sense.

Denoting by $\varphi_n(P)$, $n=1, 2, \ldots$ the eigenfunctions of the operator A and by λ_n, the corresponding eigenvalues, we have

$$(\varphi_n, \varphi_m)=\begin{cases}0, & n\neq m,\\ 1, & n=m;\end{cases} \qquad [\varphi_m, \varphi_n]=0, \quad n\neq m$$

$$A\varphi_n=\lambda_n\varphi_n.$$

Several eigenfunctions corresponding to one and the same eigenvalue, is equivalent to some of the λ_n being equal.
Multiplying the latter equation scalarly by φ_n, we further obtain

$$(A\varphi_n, \varphi_n) = \|\varphi_n\|^2 = \lambda_n. \tag{5}$$

Equation (5) proves that the functions $\varphi_n(P)$ are not normalized in energy. Putting

$$\psi_n(P) = \frac{1}{\sqrt{(\lambda_n)}} \varphi_n(P),$$

we obtain a system of functions which are orthonormal in energy and, by our assumptions, complete in energy. Corresponding to formula (2) § 12, the solution of equation (4) can be represented in the form

$$u_0(P) = \sum_{n=1}^{\infty} (f, \psi_n)\psi_n(P),$$

or, replacing $\psi_n(P)$ by its expression in terms of $\varphi_n(P)$,

$$u_0(P) = \sum_{n=1}^{\infty} \frac{(f, \varphi_n)}{\lambda_n} \varphi_n(P). \tag{6}$$

Expansion (6) into a series in the eigenfunctions of the given operator has one advantage over the general series (2) § 12. We retain in one of these series only the first N terms and we denote the finite sum thus obtained by $u_N(P)$. Since the series mentioned converges in energy, in both cases $u_N(P)$ converges to $u_0(P)$ both in the mean and in energy. However, if we substitute $u_N(P)$ in equation (4), then, generally speaking, its left-hand side will not approximate to the right; more precisely, in the general case it is impossible to assert that $Au_N \xrightarrow[N\to\infty]{\text{mean}} f(P)$ (See the author's book [11], § 8, p.41.) But if we start from series (6) and put

$$u_N(P) = \sum_{n=1}^{N} \frac{(f, \varphi_n)}{\lambda_n} \varphi_n(P)$$

then $Au_N \xrightarrow[N\to\infty]{\text{mean}} f(P)$. The proof is based upon the following statement: if the energy is defined with the aid of a positive-bounded-below operator and a certain system of functions is complete in energy, then it is also complete in the sense of convergence in the mean (as in §12, 1). Hence it follows that the system of functions $\varphi_n(P)$, $n=1, 2, \ldots$ which was assumed to be complete in energy, is also complete in the sense of convergence in the mean. Since this system is also orthonormal, any function with a finite norm may be expressed as an orthogonal series in the functions $\varphi_n(P)$; in particular

$$f(P)=\sum_{n=1}^{\infty}(f, \varphi_n)\varphi_n(P),$$

the series converging in the mean:

$$\sum_{n=1}^{N}(f, \varphi_n)\varphi_n(P)\xrightarrow[N\to\infty]{\text{mean}} f(P).$$

On the other hand

$$Au_N=\sum_{n=1}^{N}\frac{(f, \varphi_n)}{\lambda_n}A\varphi_n.$$

Since φ_n is the eigenfunction of the operator A corresponding to the eigenvalue λ_n, $A\varphi_n=\lambda_n\varphi_n(P)$ and consequently

$$Au_N=\sum_{n=1}^{N}(f, \varphi_n)\varphi_n(P);$$

hence it follows that $Au_N \xrightarrow{\text{mean}} f(P)$. Thus if we construct an approximate solution of equation (4), starting from series (6), then, for a sufficiently large N, this approximate solution satisfies equation (4) in the mean to within a prespecified degree of precision.

§ 31. Energy theorems in the eigenvalue problem

The eigenvalue problem under known conditions can be reduced to a variational problem. Let the symmetric operator A satisfy the inequality

$$(Au, u) \geq k\|u\|^2, \tag{1}$$

where k is some real number which is necessarily positive; such an operator is described as *bounded below*. In particular, any positive definite operator is bounded below since it satisfies inequality (1) for $k=0$

If the operator A is bounded below

$$\frac{(Au, u)}{(u, u)} \geq k.$$

The quantity $\dfrac{(Au, u)}{(u, u)}$ being bounded below, has a lower bound d. Obviously $d \geq k$.

We shall prove the following theorems:

THEOREM 1. *Let A be a symmetric operator bounded below and let d denote the exact lower bound of the value of the functional*

$$\frac{(Au, u)}{(u, u)}. \tag{2}$$

If there exists a function $u_0 \not\equiv 0$ such that

$$\frac{(Au_0, u_0)}{(u_0, u_0)} = d, \tag{3}$$

then d is the lowest eigenvalue of the operator A, and u_0 is the eigenfunction corresponding to this value.

Let η be an arbitrary function from the field of definition D_A of the operator A and let t be an arbitrary real number. Obviously $u_0 + t\eta \in D_A$. The function

$$\varphi(t) = \frac{(A(u_0+t\eta), u_0+t\eta)}{(u_0+t\eta, u_0+t\eta)} = \frac{t^2(A\eta, \eta) + 2t(Au_0, \eta) + (Au_0, u_0)}{t^2(\eta, \eta) + 2t(u_0, \eta) + (u_0, u_0)}$$

has a minimum at $t=0$, and $\varphi'(0)=0$.
By evaluating $\varphi'(0)$ we easily arrive at the equation

$$(u_0, u_0)(Au_0, \eta)-(Au_0, u_0)(u_0, \eta)=0,$$

or if we use equation (3),

$$(Au_0-du_0, \eta)=0. \tag{4}$$

By repeating the argument of § 11 (p.81) we find that

$$Au_0-du_0=0$$

i.e., that d and u_0 are an eigenvalue and an eigenfunction of operator A.

That d is the lowest of the eigenvalues of operator A follows from formula (1) § 30.
In fact, if λ_1 and u_1 are some eigenvalue and its corresponding eigenfunction of the operator A, then

$$\lambda_1=\frac{(Au_1, u_1)}{(u_1, u_1)}\geq \min\frac{(Au, u)}{(u, u)}=d.$$

Theorem 1 proved here reduces the problem of finding the lowest eigenvalue of a symmetric operator which is bounded below to the following variational problem of finding *the function minimizing the functional*

$$\frac{(Au, u)}{(u, u)}. \tag{5}$$

We shall indicate for this problem a further formulation which is frequently more convenient. We put $\psi=\frac{u}{\|u\|}$. Then

$$\|\psi\|=1,$$

and

$$\frac{(Au, u)}{(u, u)}=(A\psi, \psi). \tag{6}$$

Again replacing the symbol ψ by u, we can formulate our variational problem thus: *Find the minimum of the functional*

$$(Au, u) \tag{7}$$

under the supplementary condition

$$(u, u)=1. \tag{8}$$

We now state the method of determining the subsequent eigenvalues.

THEOREM 2. Let $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ be the first n consecutive eigenvalues of the symmetric operator bounded below, A, and let $u_1, u_2, \ldots, u_n$ be their corresponding orthonormal eigenfunctions. Let there exist a function $u = u_{n+1} \not\equiv 0$ which minimizes functional (5) under the supplementary conditions

$$(u, u_1) = 0, \quad (u, u_2) = 0, \quad \ldots \quad (u, u_n) = 0. \tag{9}$$

Then u_{n+1} is the eigenfunction of operator A corresponding to the eigenvalue

$$\lambda_{n+1} = \frac{(Au_{n+1}, u_{n+1})}{(u_{n+1}, u_{n+1})}. \tag{10}$$

This eigenvalue is the next following after λ_n.

Let ζ be an arbitrary function from the field of definition D_A of operator A. We put

$$\eta = \zeta - \sum_{k=1}^{n} (\zeta, u_k) u_k.$$

Then η satisfies conditions (9). In fact

$$(\eta, u_m) = (\zeta, u_m) - \sum_{k=1}^{n} (\zeta, u_k)(u_k, u_m), \quad m = 1, 2, \ldots, n. \tag{11}$$

From what was proved in § 30, the u_k are orthogonal, being eigenfunctions of a symmetric operator. Further it is assumed in the theorem that the u_k are normalized.
Therefore

$$(u_k, u_m) = \begin{cases} 0, & k \neq m \\ 1, & k = m, \end{cases}$$

and, consequently,

$$(\eta, u_m) = (\zeta, u_m) - (\zeta, u_m) = 0.$$

As well as η, the product $t\eta$ also satisfies condition (9) where t is any number, and so does the sum $u_{n+1} + t\eta$.
The function of t

$$\frac{(A(u_{n+1} + t\eta), u_{n+1} + t\eta)}{(u_{n+1} + t\eta, u_{n+1} + t\eta)}$$

is a minimum at $t=0$. By repeating the same argument as in the proof of theorem 1, we find that

$$(Au_{n+1}-\lambda_{n+1}u_{n+1},\ \eta)=0. \tag{12}$$

We consider the quantity

$$(Au_{n+1}-\lambda_{n+1}u_{n+1},\ \zeta).$$

By formulae (11) and (12)

$$\begin{aligned}(Au_{n+1}-\lambda_{n+1}u_{n+n},\ \zeta)&=(Au_{n+1}-\lambda_{n+1}u_{n+1},\ \eta)\\&\quad-\sum_{k=1}^{k}(u_k,\ \zeta)(Au_{n+1}-\lambda_{n+1}u_{n+1},u_k)\\&=-\sum_{k=1}^{n}(u_k,\ \zeta)(Au_{n+1}-\lambda_{n+1}u_{n+1},\ u_k).\end{aligned}$$

Furthermore,

$$(Au_{n+1}-\lambda_{n+1}u_{n+1},\ u_k)=(Au_{n+1},\ u_k)-\lambda_{n+1}(u_{n+1},\ u_k).$$

It is easily seen that this expression equals zero. In fact, the second term on the right vanishes by virtue of conditions (9). The first then equals

$$(Au_{n+1},\ u_k)=(u_{n+1},\ Au_k).$$

But u_k being an eigenfunction of operator A, satisfies the equation $Au_k-\lambda_k u_k=0$. Thus $Au_k=\lambda_k u_k$,

$$(Au_{n+1},\ u_k)=\lambda_k(u_{n+1},\ u_k)=0.$$

Hence it follows that $(Au_{n+1}-\lambda_{n+1}u_{n+1},\ \zeta)=0$.

Repeating the argument of § 11, we again find that $Au_{n+1}-\lambda_{n+1}u_{n+1}=0$, i.e. λ_{n+1} is an eigenvalue of the operator, and u_{n+1} the corresponding eigenfunction. It remains to prove that λ_{n+1} is the lowest eigenvalue after λ_n. Let λ' be an eigenvalue of operator A, greater than λ_n, and u' the corresponding eigenfunction.

By theorem 2 § 30, u' satisfies conditions (9). Further by formula (1) § 30,

$$\lambda'=\frac{(Au',\ u')}{(u',\ u')}\geq\lambda_{n+1},$$

since λ_{n+1} is the minimum of functional (4) under conditions (9).

As in the case of the lowest eigenvalue, the finding of λ_{n+1} can be reduced to the variational problem of *finding the minimum of function* (7) *under the supplementary conditions* (8) *and* (9).

The foregoing theorems are conditional in nature in that they give a method of constructing the eigenvalues provided that their existence is established by other means. A theorem will now be formulated which establishes the conditions for the existence of eigenvalues for a positive-bounded-below operator.

Prior to this we introduce an important concept.

We shall call an infinite set of functions *compact* in the sense of convergence of a given type if it is possible to separate from any part of it a sequence which converges to some limit in the sense of the stated type of convergence. Thus it is possible to distinguish sets of functions which are compact in the sense of convergence in energy, uniform convergence and so forth.

The concept of compactness allows us to formulate the following theorem:

THEOREM 3. *Let a positive-bounded-below operator be such that any set of functions, all of whose energy norms are bounded, is compact in the sense of convergence in the mean. Then (a) the operator has an infinite set of eigenvalues*

$$0<\lambda_1\leq\lambda_2\leq \ldots \leq\lambda_n \ldots ; \quad \lim\lambda_n=+\infty;$$

(b) the corresponding eigenfunctions form a system which is complete both in energy and in the sense of convergence in the mean.

Proof. (1) We put

$$\varphi(u)=\frac{(Au, u)}{(u, u)}=\frac{|u|^2}{\|u\|^2}, \tag{13}$$

and denote by λ_1 the exact lower bound of the functional $\varphi(u)$. Clearly $\lambda_1\geq\gamma^2$ where γ^2 is the constant in (7), § 6. We shall

prove that there exists a function u_1 satisfying the equation $\varphi(u_1)=\lambda_1$.

By the definition of the exact lower bound, no matter what the value of the positive integer n, it is possible to find a function $v_n \in D_A$ such that

$$\lambda_1 \leq \varphi(v_n) \leq \lambda_1 + \frac{1}{n}.$$

Obviously

$$\lim_{n\to\infty} \varphi(v_n) = \lambda_1.$$

The functional $\varphi(u)$ is unchanged under the multiplication of u by a constant. Multiplying the function v_n by $\|v_n\|^{-1}$ we may consider that $\|v_n\|=1$. But then $\varphi(v_n)=\|v_n\|^2$ and

$$\|v_n\|^2 \to \lambda_1. \tag{14}$$

Now let η be an arbitrary function from the field of definition D_A of operator A and let t be an arbitrary real number. Clearly

$$\frac{\|v_n+t\eta\|^2}{\|v_n+t\eta\|^2} \geq \lambda_1.$$

By replacing the square of the norm by a scalar product or an energy product, as appropriate, and using the fact that $\|v_n\|=1$, we easily reduce the latter inequality to the form

$$t^2\{\|\eta\|^2-\lambda_1\|\eta\|^2\}+2t\{[v_n,\ \eta]-\lambda_1(v_n,\ \eta)\}+\{\|v_n\|^2-\lambda_1\}\geq 0.$$

If a quadratic expression does not change sign, then its discriminant is non-positive. Hence

$$\begin{aligned} |[v_n,\ \eta]-\lambda_1(v_n,\ \eta)| &\leq \sqrt{(\|\eta\|^2-\lambda_1\|\eta\|^2)}\ \sqrt{(\|v_n\|^2-\lambda_1)} \\ &\leq \|\eta\|\ \sqrt{(\|v_n\|^2-\lambda_1)}. \end{aligned} \tag{15}$$

Let $\|\eta\|<C=\text{const}$, and let η otherwise be arbitrary. From (14) and (15) it then follows that

$$[v_n,\ \eta]-\lambda_1(u_n,\ \eta) \underset{n\to\infty}{\to} 0. \tag{16}$$

Since $\|v_n\|^2 \to \lambda_1$, then $\|v_n\|<C$, where C is some constant. Thus, if we put

$$\eta = v_n - v_m,$$

then $\|\eta\| \leq \|v_n\| + \|v_m\| < 2C$ and by virtue of inequality (16),

$$[v_n, v_n - v_m] - \lambda_1 (v_n, v_n - v_m) \underset{n \to \infty}{\to} 0.$$

Interchanging m and n and adding we obtain

$$\|v_n - v_m\|^2 - \lambda_1 \|v_n - v_m\|^2 \underset{n, m \to \infty}{\longrightarrow} 0. \tag{17}$$

The inequality $\|v_n\| < C$ means that the sequence v_n, $n = 1, 2, \ldots$ is bounded in its energy norm. By the hypotheses of the theorem this sequence is compact in the sense of convergence in the mean. We select therefrom a sequence which converges in the mean; to avoid the introduction of new symbols we denote this sequence by v_n as before. Let $v_n \overset{\text{mean}}{\longrightarrow} u_1$. Then $\|v_n - v_m\| \underset{n, m \to \infty}{\longrightarrow} 0$. From (17) it follows that $\|v_n - v_m\| \underset{n, m \to \infty}{\longrightarrow} 0$. But this means that the sequence v_n also converges to some limit in energy. It is clear that this limit also equals u_1 since $\|v_n - u_1\| \to 0$. It is also clear that $\|u_1\| = \lim \|v_n\| = 1$ and $\|u_1\| = \lim \|v_n\| = \lambda_1$ so that $\varphi(u_1) = \lambda_1$.

We now prove that $Au_1 = \lambda_1 u_1$. Proceeding to the limit as $n \to \infty$ in (16), we obtain

$$[u_1, \eta] - \lambda_1 (u_1, \eta) = 0, \qquad \eta \in D_A. \tag{18}$$

We shall find the function u which minimizes the functional

$$[u, u] - 2(u, \lambda_1 u_1); \tag{19}$$

As we know, such a function is the (possibly generalized) solution of the equation $Au_1 = \lambda_1 u_1$. From equation (18) (replacing the symbol η by u) it follows that the latter functional equals

$$[u, u] - 2[u, u_1] = \|u - u_1\|^2 - \|u\|^2.$$

It is thus clear that the function u_1 minimizes the functional (19). But then, as we have just noted, this function satisfies the equation $Au_1 = \lambda_1 u_1$. Thus it has been proved that λ_1 and u_1 are an eigenvalue and an eigenfunction of operator A. From theorem 1 it follows that λ_1 is its lowest eigenvalue.

(2) We now denote by λ_2 the exact lower bound of functional $\varphi(u)$ under the supplementary condition $(u, u_1)=0$.

Insofar as this condition restricts the class of functions on which we seek the minimum, $\lambda_2 \geq \lambda_1$. By repeating the foregoing arguments *mutatis mutandis* we find that λ_2 is the second eigenvalue of operator A and that this value corresponds to a normalized eigenfunction u_2, orthogonal to u_1. Continuing this process we construct an increasing sequence of eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \ldots$ (which, clearly, are positive) and the corresponding, orthonormal sequence of eigenfunctions $u_1(P), u_2(P), \ldots, u_n(P) \ldots$.

(3) We shall prove that $\lambda_n \to \infty$. We suppose the converse to be true and let $\lambda_n \leq C = \text{const}$. By formula (5) § 30, $\|u_n\| = \lambda_n \leq C$ i. e., the eigenfunctions are bounded in energy. By the hypothesis of the theorem, the sequence of eigenfunctions is then compact, and it is possible to separate from it a sequence u_{nk}, $k=1, 2, \ldots$ which converges in the mean. But then

$$\lim_{k, l \to \infty} \|u_{nk} - u_{nl}\|^2 = 0.$$

This, however, is impossible, since the eigenfunctions are orthonormal and thus

$$\|u_{nk} - u_{nl}\|^2 = (u_{nk} - u_{nl}, u_{nk} - u_{nl}) = \|u_{nk}\|^2 - 2(u_{nk}, u_{nl}) + \|u_{nl}\|^2 = 2.$$

(4) The system of functions u_n is complete in energy. In order to satisfy ourselves of this, we note that when the eigenfunctions are orthogonal in energy (theorem 3 § 30) λ_n can be determined as the exact lower bound of $\varphi(u)$ for the set of functions which satisfy the conditions

$$[u, u_1]=0, \quad [u, u_2]=0, \quad \ldots, \quad [u, u_{n-1}]=0.$$

If our system were incomplete in energy, then there would be functions not identically equal to zero which are orthogonal in energy to all the u_n. Denoting by $\bar{\lambda}$ the exact lower bound of $\varphi(u)$ for the functions mentioned we would find that $\bar{\lambda}$ is an eigenvalue of operator A, greater than all λ_n, and this is impossible since $\lambda_n \to \infty$.

(5) We shall carry out the proof of the fact that the system of eigenfunctions u_n is complete in the sense of convergence in the mean on the supposition that any function with a finite norm can be approximated in the mean with any degree of precision by functions with finite energy. This supposition can be proved for positive-bounded-below operators. (See the author's book [11].) On this assumption we specify a function $f(P)$ with a finite norm and select a function $g(P)$ with finite energy such that $\|f-g\|<\frac{\varepsilon}{2}$ where ε is an arbitrarily specified positive number. We approximate the function $g(P)$ by a sum of the form $a_1u_1(P)+a_2u_2(P)+\ldots+a_nu_n(P)$ such that $\left|g-(a_1u_1+a_2u_2+\ldots+a_nu_n)\right|<\frac{\varepsilon\gamma}{2}$ where γ is the constant in inequality (7) § 6. Now

$$\|f-(a_1u_1+a_2u_2+\ldots+a_nu_n)\|\leq\|f-g\|+\|g-\sum_{i=1}^{n}a_iu_i\|$$

$$\leq\|f-g\|+\frac{1}{\gamma}\left|g-\sum_{i=1}^{n}a_iu_i\right|<\varepsilon,$$

which was to be proved.

NOTE 1. If an operator possesses properties (a) and (b) of theorem 3, then it is said that it has a *discrete spectrum.*

NOTE 2. It can be proved that for an operator which is positive-bounded-below, the condition of theorem 3 is not only sufficient, but also necessary, in order that this operator have a discrete spectrum.

§ 32. Ritz's method for the eigenvalue problem

Let A be an operator which is bounded below:

$$(Au, u)\geq k\|u\|^2.$$

The finding of the eigenvalues of operator A readily reduces to finding the eigenvalues of a positive-bounded-below ope-

rator. In fact let c be any number greater than $|k|$. We re-write the equation

$$Au-\lambda u=0,$$

defining the eigenvalues of operator A, in the form

$$\tilde{A}u-\tilde{\lambda}u=0,$$

where $\tilde{A}u=Au+cu$, $\tilde{\lambda}=\lambda+c$. The operator $\tilde{A}$ is positive-bounded-below, since

$$(\tilde{A}u, u)=(Au, u)+c(u, u)\geq(c+k)\|u\|^2,$$

and $c+k>0$. If $\tilde{\lambda}$ is an eigenvalue of operator $\tilde{A}$, then $\lambda=\tilde{\lambda}-c$ is an eigenvalue of operator A and vice versa. Therefore in the sequel we shall consider that A is positive-bounded-below. We put (using the symbol "inf" for the exact lower bound)

$$d=\inf\frac{(Au, u)}{(u, u)}. \tag{1}$$

By theorem 1 § 31 d is the lowest eigenvalue of operator A, if there exists a function u_0 such that

$$d=\frac{(Au_0, u_0)}{(u_0, u_0)}.$$

If we assume that such a function exists, then the determination of the lowest eigenvalue of operator A reduces to the determination of the exact lower bound of functional (1) or, what is the same thing, the determination of the exact lower bound of the functional

$$(Au, u) \tag{2}$$

under the supplementary condition

$$(u, u)=1. \tag{3}$$

We shall show that this problem can be solved by the Ritz method. We take a sequence of coordinate functions $\varphi_n(P)$, $n=1, 2, \ldots$, subject to the three requirements: 1) the func-

tions $\varphi_n(P)$ belong to the field of definition of the operator A; 2) the system of these functions is complete in energy; 3) for any n the functions $\varphi_1(P)$, $\varphi_2(P)$, . . . , $\varphi_n(P)$, are linearly independent. We put

$$u_n(P)=\sum_{k=1}^{n}a_k\varphi_k(P),$$

where a_k are constant coefficients. We choose these coefficients such that u_n satisfies relation (3) and such that the quantity (Au_n, u_n) is a minimum. We thus require to find the minimum of a function in n variables

$$(Au_n, u_n)=\sum_{k,\,m=1}^{n}(A\varphi_k, \varphi_m)a_ka_m, \tag{4}$$

which are related by the equation

$$(u_n, u_n)=\sum_{k,\,m=1}^{n}(\varphi_k, \varphi_m)a_ka_m=1. \tag{5}$$

In order to solve this problem we use Lagrange's method of undetermined multipliers (cf. 1, **167**). We construct the function $\Phi=(Au_n, u_n)-\lambda(u_n, u_n)$ where λ is as yet an undetermined numerical factor, and we equate its partial derivatives with respect to the coefficients a_n to zero. This leads us to the system of equations

$$\sum_{k=1}^{n}a_k[(A\varphi_k, \varphi_m)-\lambda(\varphi_k, \varphi_m)]=0, \tag{6}$$
$$m=1,\ 2,\ \ldots,\ n.$$

System (6) is linear and homogeneous in the unknowns a_k which cannot all be zero simultaneously, as this would violate equation (5). Hence it follows that the determinant of system (6) must equal zero; this gives us an equation for λ:

$$\begin{vmatrix} (A\varphi_1, \varphi_1)-\lambda(\varphi_1, \varphi_1); & (A\varphi_2, \varphi_1)-\lambda(\varphi_2, \varphi_1); & \ldots & (A\varphi_n, \varphi_1)-\lambda(\varphi_n, \varphi_1) \\ (A\varphi_1, \varphi_2)-\lambda(\varphi_1, \varphi_2); & (A\varphi_2, \varphi_2)-\lambda(\varphi_2, \varphi_2); & \ldots & (A\varphi_n, \varphi_2)-\lambda(\varphi_n, \varphi_2) \\ \cdots & \cdots & \cdots & \cdots \\ (A\varphi_1, \varphi_n)-\lambda(\varphi_1, \varphi_n); & (A\varphi_2, \varphi_n)-\lambda(\varphi_2, \varphi_n); & \ldots & (A\varphi_n, \varphi_n)-\lambda(\varphi_n, \varphi_n) \end{vmatrix}=0. \tag{7}$$

If the sequence $\{\varphi_n\}$ is orthonormal, then equation (7) simplifies and assumes the form

$$\begin{vmatrix} (A\varphi_1, \varphi_1)-\lambda & (A\varphi_2, \varphi_1) & \ldots (A\varphi_n, \varphi_1) \\ (A\varphi_1, \varphi_2) & (A\varphi_2, \varphi_2)-\lambda & \ldots (A\varphi_n, \varphi_2) \\ \ldots\ldots & \ldots\ldots\ldots\ldots & \\ (A\varphi_1, \varphi_n) & (A\varphi_2, \varphi_n) & \ldots (A\varphi_n, \varphi_n)-\lambda \end{vmatrix} = 0. \qquad (8)$$

As was noted previously, the functions $\varphi_1, \varphi_2, \ldots, \varphi_n$ are linearly independent for any n. Equation (7) will then be precisely of the nth degree, since the coefficient of $(-1)^n\lambda^n$ is the Gram determinant of the functions $\varphi_1, \varphi_2, \ldots, \varphi_n$. Hence it follows that equation (7) has precisely n roots. Let λ_0 be any one of these roots. Substituting it in system (6) we make the determinant equal to zero, and the system will have non-trivial solutions. Let $a_k^{(0)}$, $k=1, 2, \ldots, n$ be such a solution. Then $\mu a_k^{(0)}$ where μ is an arbitrary numerical multiplier, also satisfies system (6). Substituting $\mu a_k^{(0)}$ in (5) we find the value of μ. Now by replacing the notation $\mu a_k^{(0)}$ by $a_k^{(0)}$ we can understand by $a_k^{(0)}$ in what follows the solution of (6) which satisfies equation (5).

Substituting $\lambda=\lambda_0$, $a_k=a_k^{(0)}$ in (6) we obtain an identity which we shall write in the form

$$\sum_{k=1}^{n} a_k^{(0)}(A\varphi_k, \varphi_m)=\lambda_0\sum_{k=1}^{n} a_k^{(0)}(\varphi_k, \varphi_m). \quad m=1, 2, \ldots, n. \qquad (9)$$

We multiply it by $a_m^{(0)}$ and sum over all m. We then obtain

$$\sum_{k, m=1}^{n} (A\varphi_k, \varphi_m)a_k^{(0)}a_m^{(0)}=\lambda_0 \sum_{k, m=1}^{n} (\varphi_k, \varphi_m)a_k^{(0)}a_m^{(0)}.$$

By virtue of equation (5), the right-hand side of the latter equation equals λ_0; its left-hand side equals $(Au_u^{(0)}, u_n^{(0)})$, where

$$u_n^{(0)}= \sum_{k=1}^{k} a_k^{(0)}\varphi_k.$$

Thus

$$\lambda=(Au_n^{(0)}, u_n^{(0)}). \qquad (10)$$

Formula (10) shows that the roots of *equation* (7) *are real if operator A is symmetric.* Further, one of the functions $u_n^{(0)}$ minimizes the quantity (4). Formula (10) now shows that *this minimum equals the least of the roots of equation* (7).

The minimum mentioned, which we shall denote below by $\lambda_n^{(0)}$ does not increase with increasing n; at the same time it is not less than the quantity d. Hence it follows that as $n \to \infty$ the quantity $\lambda_n^{(0)}$ tends to a limit which is greater than or equal to d. We now prove that this limit equals d. This will justify the application of the Ritz method to the problem of eigenvalues, at least for the case which we are concerned with, the determination of the lowest eigenvalue. We introduce the energy product and the norm of the energy, putting, as usual,

$$[u, v] = (Au, v); \qquad |u|^2 = (Au, u).$$

For any $\varepsilon > 0$ there exists by the definition of exact lower bound a function $u' \in D_A$ such that $(u', u') = 1$ and

$$d \leq (Au', u') < d + \varepsilon,$$

or, what is the same thing,

$$\sqrt{(d)} \leq |u'| < \sqrt{(d+\varepsilon)}.$$

Since the sequence $\varphi_n(P)$, $n = 1, 2, \ldots$, is complete in energy, it is possible to find a function u'_N of the form

$$u'_N = \sum_{k=1}^{N} b_k \varphi_k, \quad b_k = \text{const},$$

such that $|u' - u'_N| < \sqrt{\varepsilon}$.

Hence

$$|u'_N| \leq |u'| + \sqrt{\varepsilon} < \sqrt{(d+\varepsilon)} + \sqrt{\varepsilon}$$

or

$$(Au'_N, u'_N) < (\sqrt{(d+\varepsilon)} + \sqrt{\varepsilon})^2.$$

From (1) it follows that $\|u'_N - u'\| \leq \frac{1}{\sqrt{d}} |u'_N - u'| < \sqrt{\frac{\varepsilon}{d}}$.

Hence $\|u'_N\| \geq \|u'\| - \sqrt{\frac{\varepsilon}{d}} = 1 - \sqrt{\frac{\varepsilon}{d}}$ and, consequently

$$d \leq \frac{(Au'_N, u'_N)}{(u'_N, u'_N)} = \frac{|u'_N|^2}{\|u'_N\|^2} < \frac{(\sqrt{(d+\varepsilon)} + \sqrt{\varepsilon})^2}{\left(1 - \sqrt{\frac{\varepsilon}{d}}\right)^2} = d + \eta,$$

where η tends to zero with ε. Further $\lambda_N^{(0)}$ is the minimum of the expression of the form

$$\frac{(Au_N, u_N)}{(u_N, u_N)}; \quad u_N = \sum_{k=1}^{N} a_k \varphi_k,$$

thus, in particular,

$$d \leq \lambda_N^{(0)} \leq \frac{(Au'_N, u'_N)}{(u'_N, u'_N)} < d + \eta.$$

If $n \geq N$ then $\lambda_n^{(0)} \leq \lambda_N^{(0)}$ and, consequently, $d \leq \lambda_n^{(0)} < d + \eta$. The latter inequality shows that

$$\lim_{n \to \infty} \lambda_n^{(0)} = d, \tag{11}$$

which was to be proved.

We now pass on to the determination of the subsequent eigenvalues. To obtain an approximate value of the second eigenvalue we seek the minimum of the scalar product (4) under the supplementary conditions

$$(u_n, u_n) = 1 \tag{5}$$

and

$$(u_n^{(0)}, u_n) = \sum_{k,m=1}^{n} (\varphi_k, \varphi_m) a_k^{(0)} a_m = 0, \tag{12}$$

where $u_n^{(0)} = \sum_{k=1}^{n} a_k^{(0)} \varphi_k$ is the approximate value of the first normalized eigenfunction of operator A.

By Lagrange's method we construct the expression

$$(Au_n, u_n) - \lambda(u_n, u_n) - 2\mu(u_n, u_n^{(0)})$$

and equate its partial derivatives with respect to the a_k to zero. This leads us to the system

$$\sum_{k=1}^{n} \{a_k[(A\varphi_k, \varphi_m) - \lambda(\varphi_k, \varphi_m)] - \mu(\varphi_k, \varphi_m) a_k^{(0)}\} = 0. \tag{13}$$

We multiply both sides of (13) by $a_m^{(0)}$ and sum over m:

$$\sum_{k,\,m=1}^{n} a_k a_m^{(0)}[(A\varphi_k,\,\varphi_m)-\lambda(\varphi_k,\,\varphi_m)] - \mu \sum_{k,\,m=1}^{n} a_k^{(0)} a_m^{(0)}(\varphi_k,\,\varphi_m)=0. \tag{14}$$

The second summation, as is easily seen, is $(u_n^{(0)},\,u_n^{(0)})=1$. In the first summation we replace the index m by k and vice versa and sum firstly over k, then over m. We then reduce the first summation to the form

$$\sum_{m=1}^{n} a_m \sum_{k=1}^{n} a_k^{(0)}[(A\varphi_m,\,\varphi_k)-\lambda(\varphi_m,\,\varphi_k)]. \tag{15}$$

The internal sum equals

$$\sum_{k=1}^{n} a_k^{(0)}[(\varphi_k,\,A\varphi_m)-\lambda(\varphi_k,\,\varphi_m)]$$

or since operator A is symmetric,

$$\sum_{k=1}^{n} a_k^{(0)}[(A\varphi_k,\,\varphi_m)-\lambda(\varphi_k,\,\varphi_m)].$$

By virtue of equations (6) which the numbers $a_k^{(0)}$ satisfy for $\lambda=\lambda_n^{(0)}$ the latter expression equals

$$(\lambda_n^{(0)}-\lambda)\sum_{k=1}^{n} a_k^{(0)}(\varphi_k,\,\varphi_m)=(\lambda_n^{(0)}-\lambda)\left(\sum_{k=1}^{n} a_k^{(0)}\varphi_k,\,\varphi_m\right)$$
$$=(\lambda_n^{(0)}-\lambda)(u_n^{(0)},\,\varphi_m)=(\lambda_n^{(0)}-\lambda)(\varphi_m,\,u_n^{(0)}).$$

Expression (15) assumes the form

$$(\lambda_n^{(0)}-\lambda)\sum_{m=1}^{n} a_m(\varphi_m,\,u_n^{(0)})=(\lambda_n^{(0)}-\lambda)(u_n,\,u_n^{(0)}),$$

which equals zero, because of (12). Now from (14) it follows that $\mu=0$ and the system (13) is identical with the system (6). Hence, as above, we conclude that the minimum sought equals λ, which is a root of equation (7). It is easily seen that

this time we must take the root second in value of the equation mentioned.

Similarly, it is possible to construct approximate values of higher eigenvalues; they are all roots of equation (7).

The question of the convergence of the approximate values of the eigenvalues to their exact values will be investigated later (§ 78).

§ 33. Another form of Ritz's method; the case of natural boundary conditions

We shall start from a positive-bounded-below operator with a discrete spectrum. From theorem 1 § 31 its lowest eigenvalue λ_1 equals

$$\lambda_1 = \inf \frac{(Au, u)}{(u, u)}.$$

The exact lower bound is taken with respect to the set of all the functions which constitute the field of definition of operator A, so that the previous formula can thus be written in a clearer form.

$$\lambda_1 = \inf_{u \in D_A} \frac{(Au, u)}{(u, u)}, \tag{1}$$

which can further be put in the form

$$\lambda_1 = \inf_{u \in D_A} \frac{|u|^2}{\|u\|^2}. \tag{2}$$

The ratio in formula (1) is meaningful only for functions from the field of definition of operator A, but in formula (2) the ratio can be defined for the wider class of functions with finite energy. We shall prove that in this case the exact lower bound of the ratio mentioned is unchanged, so that, if we denote the set of functions with finite energy by H_A, then

$$\lambda_1 = \inf_{u \in H_A} \frac{|u|^2}{\|u\|^2}. \tag{3}$$

In extending the class of functions, the exact lower bound can only decrease, therefore in each case

$$\inf_{u\in H_A}\frac{|u|^2}{\|u\|^2}\leq\lambda_1.$$

For brevity we use the notation

$$\delta=\inf_{u\in H_A}\frac{|u|^2}{\|u\|^2}$$

and suppose that $\delta<\lambda_1$. By the definition of exact lower bound, it is possible to find, for any number $\varepsilon>0$ (however small) a function $\bar{u}(P)\in H_A$ such that

$$\delta\leq\frac{|\bar{u}|^2}{\|\bar{u}\|^2}<\delta+\varepsilon.$$

The ratio $|u|^2/\|u\|^2$ is unchanged if u is multiplied by any constant. Thus we can assume that $\|\bar{u}\|=1$, and then

$$\delta\leq|u|^2<\delta+\varepsilon.$$

On the other hand, by the definition of functions with finite energy, there exists a function $u'\in D_A$ such that $|u'-\bar{u}|<\varepsilon$ and, consequently $\|u'-u\|<\varepsilon/\gamma$ where γ is the constant of inequality (7) § 6.

We now estimate the ratio

$$\frac{|u'|^2}{\|u'\|^2}.$$

By the triangle inequality we have

$$\frac{|u'|^2}{\|u'\|^2}\leq\left\{\frac{|\bar{u}|+|u'-\bar{u}|}{\|\bar{u}\|-\|u'-\bar{u}\|}\right\}^2<\left\{\frac{\sqrt{(\delta+\varepsilon)}+\sqrt{\varepsilon}}{1-\sqrt{(\varepsilon/\gamma)}}\right\}^2<\delta+\eta,$$

where η tends to zero with ε. By choosing an ε so small that $\delta+\eta<\lambda_1$ we arrive at a contradiction of formula (2). But then the inequality $\delta<\lambda_1$ is impossible and equation (3) is proved. It is clear that the exact lower bound in (3) is attained if $u=u_1$ where u_1 is the eigenfunction of operator A corresponding to the eigenvalue λ_1.

The Ritz method can be also developed for finding the minimum of the ratio $\frac{|u|^2}{\|u\|^2}$, $u \in H_A$. For this we select a sequence of coordinate functions $\varphi_n(P)$, $u=1, 2$, subject to conditions 2) and 3) § 32; we replace condition 1) § 32 by the following: 1a) the functions $\varphi_n(P)$ have finite energy. As in § 32 we put

$$u_n(P)=\sum_{k=1}^{n} a_k \varphi_k(P),$$

and we require that

$$\|u_n\|^2=\sum_{k,\,m=1}^{n} a_k a_m(\varphi_k,\ \varphi_m)=1$$

and that

$$|u_n|^2=\sum_{k,\,m=1}^{n} a_k a_m[\varphi_k,\ \varphi_m]$$

should be a minimum.

Repeating the reasoning of § 32, we arrive at the equation

$$\begin{vmatrix} [\varphi_1, \varphi_1]-\lambda(\varphi_1, \varphi_1); [\varphi_2, \varphi_1]-\lambda(\varphi_2, \varphi_1); \ldots; [\varphi_n, \varphi_1]-\lambda(\varphi_n, \varphi_1) \\ [\varphi_1, \varphi_2]-\lambda(\varphi_1, \varphi_2); [\varphi_2, \varphi_2]-\lambda(\varphi_2, \varphi_2); \ldots; [\varphi_n, \varphi_2]-\lambda(\varphi_n, \varphi_2) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ [\varphi_1, \varphi_n]-\lambda(\varphi_1, \varphi_n); [\varphi_2, \varphi_n]-\lambda(\varphi_2, \varphi_n); \ldots; [\varphi_n, \varphi_n]-\lambda(\varphi_n, \varphi_n) \end{vmatrix} = 0. \quad (4)$$

We note that equation (4) is equivalent to equation (7) § 32, if we select the coordinate functions from the field of definition D_A of operator A.

As in the previous section it can be proved that the roots of equation (4) give the approximate values of the eigenvalues of the operator, and that as $n \to \infty$ the lowest root of equation (4) tends to the lowest eigenvalue of the operator.

Form (4) of the equation for the approximate values of the eigenvalues is of special importance in the case where the functions entering into the field of definition D_A of the

operator also satisfy, among others, natural boundary conditions. As we already know (§ 17) functions with finite energy need not satisfy these conditions. Hence follows a result of great practical importance: if the equation for the approximate determination of the eigenvalues is written in the form (4) and not in the form (7) § 32, then it is unnecessary to take into account the fact that the coordinate functions satisfy the natural boundary conditions.

§ 34. Equations of the form $Au-\lambda Bu=0$

In considering the equation

$$Au-\lambda Bu=0, \tag{1}$$

we shall assume that both the operators A and B are positive-bounded-below and that D_A constitutes a part of D_B. For any function entering into D_A it is possible to define the energy in two ways, by connecting it either with operator A or with operator B. We shall distinguish these two forms of the energy by speaking of "the energy of operator A" or "the energy of operator B".

In the present section we shall formulate a series of theorems relating to the eigenvalues and eigenfunctions of equation (1). We shall not present their proofs, because they resemble almost word for word the analogous proofs for the equation $Au-\lambda u=0$.

THEOREM 1. *The eigenvalues of equation* (1) *are real.*

We note that if λ_0 and u_0 are an eigenvalue and an eigenfunction of equation (1), then

$$\lambda_0=\frac{(Au_0, u_0)}{(Bu_0, u_0)}=\frac{\|u_0\|_A^2}{\|u_0\|_B^2}. \tag{2}$$

Hence it follows, *inter alia*, that the eigenvalues of equation (1) are not only real but positive as well.

THEOREM 2. *The eigenfunctions of equation* (1) *which correspond to distinct eigenvalues are orthogonal in the energy of operator B.*

Thus if λ_1 and λ_2 are distinct eigenvalues of equation (1) and u_1 and u_2 are the corresponding eigenfunctions then

$$(Bu_1, u_2) = [u_1, u_2]_B = 0. \tag{3}$$

Hence, as in § 30, it follows that *the set of all the eigenfunctions of equation* (1) *can be considered as orthonormal in the energy of the operator B.* In this case there occurs:

THEOREM 3. *The system of eigenfunctions of equation* (1) *is orthogonal in the energy of operator A.*

THEOREM 4. *Let the operators A and B be such that any set of functions, for which all the norms of the energy of operator A are bounded is compact in the sense of the norm of the energy of operator B. Then a) equation* (1) *has an infinite set of eigenvalues*

$$0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \ldots,$$

and $\lambda_n \to \infty$ *as* $n \to \infty$; *b) the corresponding eigenfunctions constitute a system which is complete both in the energy of operator A and in the energy of operator B.*

THEOREM 5. *Let d be the lower bound of the functional*

$$\frac{(Au, u).}{(Bu, u)}. \tag{4}$$

If there exists a function u_0 *such that*

$$\frac{(Au_0, u_0)}{(Bu_0, u_0)} = d,$$

then d is the lowest eigenvalue of equation (1) *and* u_0 *is the corresponding eigenfunction of this equation.*

THEOREM 6. *Let* $\lambda_1, \lambda_2, \ldots, \lambda_n$ *be (in order of increasing value) the first n eigenvalues of equation* (1), *and let* $u_1(P), u_2(P), \ldots, u_n(P)$, *be the corresponding eigenfunctions. Let there exist a function* $u_{n+1}(P)$ *which minimizes the functional* (4) *under the supplementary conditions*

$$(Bu, u_1) = (Bu, u_2) = \ldots = (Bu, u_n) = 0. \tag{5}$$

Then u_{n+1} *is the eigenfunction of equation* (1) *corresponding to the eigenvalue*

$$\lambda_{n+1}=\frac{(Au_{n+1},\ u_{n+1})}{(Bu_{n+1},\ u_{n+1})}. \tag{6}$$

This eigenvalue is the next following after λ_n.

Theorem 5 reduces the problem of finding the lowest eigenvalue of equation (1) to the variational problem of *finding the minimum of the functional* $(Au,\ u)$ *under the supplementary condition*

$$(Bu,\ u)=1. \tag{7}$$

Similarly finding the $(n+1)$th eigenvalue reduces to the problem *of finding the minimum of the functional* $(Au,\ u)$ *under the supplementary conditions* (7) *and* (5).

We shall not dwell on the details of the application of the Ritz method to equation (1). We shall merely state that the coordinate functions must be subjected to three conditions of which conditions 1) and 3) are the same as in § 32 and condition 2) is as follows: the system of coordinate functions is complete in the energy of operator A. Equations (6) and (7) § 32 are then replaced by the equations

$$\sum_{k=1}^{n} a_k[(Au_k,\ u_m)-\lambda(Bu_k,\ u_m)]=0, \quad m=1,\ 2,\ \ldots,\ n; \tag{8}$$

$$\begin{vmatrix} (Au_1,u_1)-\lambda(Bu_1,u_1); & (Au_2,u_1)-\lambda(Bu_2,u_1); & \ldots; & (Au_n,u_1)-\lambda(Bu_n,u_1) \\ (Au_1,u_2)-\lambda(Bu_1,u_2); & (Au_2,u_2)-\lambda(Bu_2,u_2); & \ldots; & (Au_n,u_2)-\lambda(Bu_n,u_2) \\ \cdot & \cdot & \cdot & \cdot \\ (Au_1,u_n)-\lambda(Bu_1,u_n); & (Au_2,u_n)-\lambda(Bu_2,u_n); & \ldots; & (Au_n,u_n)-\lambda(Bu_n,u_n) \end{vmatrix}=0. \tag{9}$$

Condition 2) can be relaxed by merely requiring that the coordinate functions possess finite energy for operator A. It can be then proved that they also possess finite energy for operator B. Instead of equations (8) and (9) we must now write

$$\sum_{k=1}^{n} a_k\{[u_k,\ u_m]_A-\lambda[u_k,\ u_m]_B\}=0, \quad m=1,\ 2,\ \ldots,\ n \tag{8_1}$$

and

$$\begin{vmatrix} [u_1, u_1]_A - \lambda[u_1, u_1]_B ; [u_2, u_1]_A - \lambda[u_2, u_1]_B ; \ldots ; [u_n, u_1]_A - \lambda[u_n, u_1]_B \\ [u_1, u_2]_A - \lambda[u_1, u_2]_B ; [u_2, u_2]_A - \lambda[u_2, u_2]_B ; \ldots ; [u_n, u_2]_A - \lambda[u_n, u_2]_B \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ [u_1, u_n]_A - \lambda[u_1, u_n]_B ; [u_2, u_n]_A - \lambda[u_2, u_n]_B ; \ldots ; [u_n, u_n]_A - \lambda[u_n, u_n]_B \end{vmatrix} = 0. \tag{9_1}$$

The observation about natural conditions (§ 33) still holds true.

§ 35. The eigenvalues of an ordinary differential equation

In the present section we shall investigate the question of the eigenvalues of an ordinary differential equation, considering in greater detail the case of a second order equation under various types of boundary conditions. For equations of order higher than the second we shall restrict ourselves to the simplest type of boundary conditions (see below), formula (28); boundary conditions of other types will be considered in the next section for the particular case of the stability of a rod under longitudinal compression.

1. We consider the second-order differential operator

$$Au = -\frac{\mathrm{d}}{\mathrm{d}x}\left[p(x)\frac{\mathrm{d}u}{\mathrm{d}x}\right] + r(x)u, \tag{1}$$

where the function $u(x)$ is specified on the segment $a \leq x \leq b$.

We shall assume that the coefficients $p(x)$ and $r(x)$ are subject to the conditions of § 20, in particular,

$$p(x) \geq 0,\ r(x) \geq 0, \quad A = \int_a^b \frac{\mathrm{d}x}{p(x)} < \infty, \tag{2}$$

and that $u(x)$ satisfies (see § 20) boundary conditions which guarantee that operator (1) is positive-bounded-below. We

pose the problem of the eigenvalues and eigenfunctions of operator A.

2. First of all we investigate the simpler case where at one of the ends of the segment $a \leq x \leq b$, say at the left-hand end, the boundary condition

$$u(a) = 0, \tag{3}$$

is satisfied, and at the right-hand end the condition is of the general form

$$\gamma u'(b) + \delta u(b) = 0, \tag{4}$$

where $\gamma \geq 0$, $\delta \geq 0$ and at least one of the quantities γ and δ is different from zero.

In order to solve our problem it will be useful initially to make an estimate of the energy defined by operator A. We have (see formula (5) § 20)

$$\|u\|^2 = \int_a^b (pu'^2 + ru^2)\,dx + \frac{\delta}{\gamma} u^2(b).$$

If $\gamma = 0$, the last term does not exist. In each case it is non-negative. Dropping the last term and also the non-negative term under the integral sign, we obtain the estimate sought

$$\|u\|^2 \geq \int_a^b p(x) u'^2(x)\,dx. \tag{5}$$

The second step is connected with the use of the following theorem from the theory of integral equations (see, e.g. the author's book [7] theorem 2 § 20).

THEOREM 1. *If*

$$u(x) = \int_a^b K(x, t) v(t)\,dt \tag{6}$$

and the double integral

$$\int_a^b \int_a^b K^2(x, t)\,dx\,dt \tag{7}$$

is finite, then the integral operator (6) *transforms any set of functions, each of whose norms is bounded, into a set which is compact in the sense of convergence in the mean.*

Thus if M is a set of functions and for each function $v(x) \in M$ the inequality

$$\|v\| = \left\{ \int_a^b v^2(x)\,dx \right\}^{\frac{1}{2}} \leq C, \quad C = \text{const},$$

is valid, then it is possible to separate from this set a sequence $v_n(x)$ such that the sequence generated from it by formula (6)

$$u_n(x) = \int_a^b K(x, t) v_n(t)\,dt$$

converges in the mean to some limit.

Let some function $u(x)$ satisfy condition (3) and have a continuous derivative. Then, clearly

$$u(x) = \int_a^x u'(t)\,dt.$$

By putting

$$v(x) = \sqrt{[p(x)]}u'(x), \quad K_0(x, t) = \begin{cases} \dfrac{1}{\sqrt{p(t)}}, & a \leq t \leq x. \\ 0, & x \leq t \leq b, \end{cases} \tag{8}$$

we can represent the last equation in the form

$$u(x) = \int_a^b K_0(x, t) v(t)\,dt. \tag{6$_1$}$$

This identity was obtained on the assumption that $u(x)$ has a continuous derivative but it can be shown to hold also for arbitrary functions with finite energy.

It is easily seen that in our case integral (7) is finite so that theorem 1 is applicable. In fact, from formula (8) it is clear that for $a \leq x$, $t \leq b$

$$0 \leq K_0(x, t) \leq \frac{1}{\sqrt{p(t)}},$$

and so

$$\int_a^b \int_a^b K_0^2(x, t)\mathrm{d}x\mathrm{d}t \leq \int_a^b \int_a^b \frac{\mathrm{d}x\mathrm{d}t}{p(t)} = A(b-a) < \infty.$$

Now let a set of functions $u(x)$ be given each of whose energy norms is bounded. By formula (5)

$$\int_a^b p(x)u'^2(x)\mathrm{d}x \leq C^2,$$

or, if we introduce the functions $v(x)$ given by formula (8), $\|v\| \leq C$. By theorem 1 of the present section the set of functions $u(x)$ is compact and the condition of theorem 3 § 31 is satisfied. Hence it follows that *under boundary conditions* (2) *and* (3) *operator* (1) *has a discrete spectrum.*

We now formulate this result at greater length:
there exists an infinitely increasing sequence of positive numbers $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \leq \ldots$, with corresponding functions $u_1(x)$, $u_2(x), \ldots, u_n(x), \ldots$ which satisfy the differential equation

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(p\frac{\mathrm{d}u_n}{\mathrm{d}x}\right) + ru_n - \lambda u_n = 0 \tag{9}$$

and boundary conditions (2) and (3). The functions $u_n(x)$ are orthonormal in the ordinary sense:

$$(u_n, u_m) = \int_a^b u_n(x)u_m(x)\mathrm{d}x = \begin{cases} 0, & n \neq m \\ 1, & n = m \end{cases} \tag{10}$$

and orthogonal in energy

$$[u_n, u_m] = \int_a^b (pu_n'u_m' + ru_nu_m)\mathrm{d}x + \frac{\delta}{\gamma}u_n(b)u_m(b) = 0, \quad n \neq m. \tag{11}$$

Moreover

$$\|u_n\|^2 = \int_a^b (pu_n'^2 + ru_n^2)\mathrm{d}x + \frac{\delta}{\gamma}u_n^2(b) = \lambda_n. \tag{12}$$

The system of functions $u_n(x)$, $n=1, 2, \ldots$ is complete both in energy and in the sense of convergence in the mean.

It may happen that the function $r(x)$ also assumes negative values. Suppose that this function still remains bounded: $|r(x)| \leq c$. Then the operator $A_c u = Au + (c+1)u$ satisfied the conditions of this section. And there follows the existence of an infinite sequence of positive eigenvalues $\lambda_1, \lambda_2, \ldots$ for the operator A_c. But in this case operator A has an infinite set of eigenvalues $\lambda_1 - c - 1, \lambda_2 - c - 1, \ldots, \lambda_n - c - 1, \ldots$, and a finite number of them can be negative or equal to zero.

3. Consider operator (1) in connexion with functions which satisfy boundary conditions of the general form

$$\alpha u'(a) - \beta u(a) = 0, \quad \gamma u'(b) + \delta u(b) = 0, \tag{13}$$

where $\alpha > 0$, $\lambda > 0$, $\beta \geq 0$, $\delta \geq 0$ and at least one of the coefficients β and δ is different from zero. For the sake of argument let $\beta > 0$. We shall prove that in this case also the conditions of theorem 1 § 32 are satisfied so that operator (1) under boundary conditions (13) has a discrete spectrum.

By formula (5) § 20, in the given case

$$\|u\|^2 = \int_a^b (pu'^2 + ru^2)\,dx + \frac{\beta}{\alpha} p(a)u^2(a) + \frac{\delta}{\gamma} p(b)u^2(b).$$

Dropping the non-negative term ru^2 in the integral and also the non-negative term $\frac{\delta}{\gamma} p(b)u^2(b)$ outside it, we find that

$$\|u^2\| \geq \int_a^b pu'^2\,dx + \frac{\beta}{\alpha} p(a)u^2(a). \tag{14}$$

Let there be given a set M of functions, each of whose energy norms is bounded.

$$\|u\| \leq C = \text{const}, \quad u \in M.$$

From inequality (14) follows

$$\int_a^b pu'^2 \mathrm{d}x \leq C^2, \tag{15}$$

$$\sqrt{p(a)}|u(a)| \leq C\sqrt{\frac{\alpha}{\beta}}. \tag{16}$$

We note the identity

$$u(x) = u(a) + \int_a^x u'(t)\mathrm{d}t;$$

which, if we use the notation of (8), can be represented in the form

$$u(x) = u(a) + \int_a^b K_0(x, t)v(t)\mathrm{d}t. \tag{17}$$

By virtue of inequality (15) $\|v\| \leq C$ and it is possible to select from the set M a sequence $u_n(x)$ such that the functions

$$\int_a^b K_0(x, t)v_n(t)\mathrm{d}t, \quad v_n(t) = \sqrt{[p(t)]}u_n(t) \tag{18}$$

converge in the mean. Furthermore by virtue of the inequality (16) $|u_n(a)|$ is bounded. By the Bolzano–Weierstrass lemma (see G. M. Fikhtengol'ts [1], p 105) it is possible to select from the sequence $u_n(x)$, $n = 1, 2, \ldots$ a partial sequence $u_{n_k}(x)$, $k = 1, 2, \ldots$, such that the sequence $u_{n_k}(a)$ converges. Now from equation (17) it follows that the sequence $u_{n_k}(x)$ converges in the mean.

4. We now consider operator (1) under the boundary conditions

$$u'(a) = u'(b) = 0. \tag{19}$$

We can always consider that $r(x) \geq 1$ – as for this it suffices to replace the operator Au by $Au + cu$, where c is a sufficiently large positive constant. Thus all the eigenvalues of

operator A are simply increased by c. Under the boundary conditions (19) we have

$$\|u\|^2=\int_a^b (pu'^2+ru^2)\mathrm{d}x. \tag{20}$$

Let a set M of functions, each of whose energy norms is bounded, be given:

$$\|u\|\leq C, \quad u\in M.$$

Then, as follows from (20),

$$\int_a^b pu'^2\mathrm{d}x\leq C^2 \tag{21}$$

and since $r\geq 1$

$$\int_a^b u^2\mathrm{d}x\leq C^2. \tag{22}$$

By putting $v(x)=\sqrt{[\mathrm{p}(x)]}u'(x)$ we again arrive at equation (17). As before we satisfy ourselves, on the basis of the inequality (21), of the possibility of selecting a sequence of functions $u_n(x)$ such that the sequence (18) converges in the mean. We now prove that $u_n(a)$ is bounded. Denoting the integral in (17) by $w_n(x)$, we have $u_n(a)=u_n(x)-w_n(x)$. Hence

$$u_n^2(a)\leq 2u_n^2(x)+2w_n^2(x).$$

Integrating this inequality over the range from a to b, we find

$$u_n^2(a)\leq \frac{2}{b-a}\int_a^b u_n^2(x)\mathrm{d}x+\frac{2}{b-a}\int_a^b w_n^2(x)\mathrm{d}x. \tag{23}$$

By virtue of the inequality (22) the first integral in (23) does not exceed C^2 and we shall prove that the second integral in (23) is also bounded. In fact, by the Cauchy–Bunyakovsky inequality

$$w_n^2(x)\leq \int_a^b K_0^2(x,\,t)\mathrm{d}t\int_a^b v_n^2(t)\mathrm{d}t.$$

By the inequality (21),

$$\int_a^b v_n^2(t)\,dt = \int_a^b p(t)u_n'^2(t)\,dt \leq C^2.$$

Further, by formula (8) $0 \leq K_0(x, t) \leq \frac{1}{[\sqrt{p(t)}]}$, and thus

$$\int_a^b K_0^2(x, t)\,dt \leq \int_a^b \frac{dt}{p(t)} = A.$$

Therefore $w_n^2(x) \leq AC^2$ and, consequently,

$$\int_a^b w_n^2(x)\,dx \leq AC^2(b-a).$$

Now

$$u_n^2(a) \leq \frac{2C^2}{b-a} + 2AC^2,$$

and $|u_n(a)|$ is bounded.

Again applying the Bolzano–Weierstrass theorem and repeating the reasoning at the end of subsection 3, we satisfy ourselves that it is possible to separate from the set M a sequence which converges in the mean. Hence on the basis of theorem 3 § 30 it follows that under conditions (19) operator (1) has a discrete spectrum.

5. The conclusions of the previous sub-sections of the present section can easily be extended for the equation

$$-\frac{d}{dx}\left(p\frac{du}{dx}\right) + r(x)u - \lambda\varrho(x) = 0, \tag{24}$$

in which $p(x)$ and $r(x)$ satisfy the conditions of 1 and $\varrho(x)$ satisfies the inequalities $\varrho_0 \leq \varrho(x) \leq \varrho_1$, where ϱ_0, ϱ_1, are certain positive constants. For the sake of argument we shall consider the simplest of boundary conditions

$$u(a) = u(b) = 0, \tag{25}$$

although the consideration of conditions of other types presents no difficulties. Equation (24) belongs to the type studied in

§ 34; in fact the operator $Bu = \varrho(x)u$ is positive-bounded-below, since

$$(Bu, u) = \int_a^b \varrho(x)u^2(x)\mathrm{d}x \geq \varrho_0 \int_a^b u^2(x)\mathrm{d}x = \varrho_0 \|u\|^2.$$

In order to prove that equation (24) has a discrete spectrum it suffices to satisfy ourselves that the conditions of theorem 4 § 34 are obeyed. In our case

$$\|u\|_A^2 = \int_a^b (pu'^2 + ru^2)\mathrm{d}x; \quad \|u\|_B^2 = \int_a^b \varrho u^2 \mathrm{d}x. \tag{26}$$

From identity (6_1) follows

$$\sqrt{[r(x)]}u(x) = \int_a^b K_1(x, t)v(t)\mathrm{d}t; \quad K_1(x, t) = \sqrt{[\varrho(x)]}K_0(x, t).$$

Because the function $\varrho(t)$ is bounded, the integral

$$\int_a^b \int_a^b K_1^2(x, t)\mathrm{d}x\mathrm{d}t = \int_a^b \int_a^b \varrho(x)K^2(x, t)\mathrm{d}x\mathrm{d}t$$

is finite and we can apply theorem 1 of the present section. From each set of functions $v(x)$ with bounded norms it is possible to separate a sequence $v_n(x)$ such that the sequence of corresponding functions $u_n(x)\sqrt{[\varrho(x)]}$ converges in the mean, or, what is the same thing, the sequence $u_n(x)$ converges in the energy of operator B. Now let a set of functions $u(x)$ for which $\|u\|_A \leq C$ be given. Then, as is seen from (26),

$$\|v\| = \left\{\int_a^b p(x)u'^2(x)\mathrm{d}x\right\}^{\frac{1}{2}} \leq C$$

and, as has just been proved, it is possible to select from the given set of functions $u(x)$ a sequence which converges in the energy of the operator B. Thus it has been proved that equation (24) has a discrete spectrum, i.e. that there exists a sequence of functions $u_n(x)$, not identically equal to zero and obviously corresponding to them, the positive numbers

λ_n, satisfying equation (24) and boundary conditions (25); the functions $u_n(x)$ are orthonormal in the energy of operator B (i.e. orthonormal with weight $\varrho(x)$ (see § 10).

$$[u_n, u_m]_B = \int_a^b \varrho(x) u_n(x) u_m(x)\,dx = \begin{cases} 0, & m \neq n, \\ 1, & m = n, \end{cases}$$

and orthogonal in the energy of operator A. In this case

$$[u_n, u_m]_A = \int_a^b (p u_m' u_n' + r u_m u_n)\,dx = \begin{cases} 0, & m \neq n, \\ \lambda_n, & m = n. \end{cases}$$

We have proved the discrete nature of the spectrum of operator (1) [and also of equation (24)] on the assumption that the integral

$$\int_a^b \frac{dx}{p(x)}$$

is finite. In point of fact it is also possible to prove the discrete nature of the spectrum on the basis of weaker assumptions. Thus, it suffices to assume that the integral

$$\int_a^b \frac{(x-a)dx}{p(x)},$$

remains finite, in which case in certain instances it is necessary to change the nature of the boundary conditions at the point $x=a$. For the details of this see, e.g. the author's paper [15].

6. We now turn to the equation of arbitrary even order:

$$Au - \lambda Bu = \sum_{k=0}^{m} (-1)^k \frac{d^k}{dx^k}\left[p_k(x)\frac{d^k u}{dx^k}\right]$$

$$-\lambda \sum_{k=0}^{s} (-1)^k \frac{d^k}{dx^k}\left[\varrho_k(x)\frac{d^k u}{dx^k}\right] = 0, \qquad (27)$$

in which $s<m$. We shall consider this equation under the very simple boundary conditions

$$u(a)=u'(a)=\ldots=u^{(m-1)}(a)$$
$$=u(b)=u'(b)=\ldots=u^{(m-1)}(b)=0. \tag{28}$$

In order to simplify the reasoning we shall assume that all the coefficients $p_k(x)$ and $\varrho_k(x)$ are non-negative and bounded and that $p_m(x)\geq p_0$, $\varrho_s(x)\geq\varrho_0$ where p_0 and ϱ_0 are positive constants. With these assumptions the operators A and B in equation (27) are positive-bounded-below (see § 20) and

$$\|u\|_A^2=\int_a^b\sum_{k=0}^{m}p_k(x)\left(\frac{\mathrm{d}^k u}{\mathrm{d}x^k}\right)^2\mathrm{d}x, \tag{29}$$

$$\|u\|_B^2=\int_a^b\sum_{k=0}^{s}\varrho_k(x)\left(\frac{\mathrm{d}^k u}{\mathrm{d}x^k}\right)^2\mathrm{d}x. \tag{30}$$

We note further that convergence in energy of operator B implies convergence in the mean both of the functions themselves and of their derivatives of order up to s inclusive.

We shall prove that, under the conditions indicated, equation (27) has a discrete spectrum. To this end it suffices to establish that the conditions of theorem 4 § 34 are satisfied.

a) Let there be given a set M of functions which is bounded in the energy of operator A: $\|u\|_A\leq C$, $u\in M$ or, at greater length,

$$\int_a^b\sum_{k=0}^{m}p_k(x)\left(\frac{\mathrm{d}^k u}{\mathrm{d}x^k}\right)^2\mathrm{d}x\leq C^2.$$

Then

$$\int_a^b\left(\frac{\mathrm{d}^m u}{\mathrm{d}x^m}\right)^2\mathrm{d}x\leq\frac{C^2}{p_0}. \tag{31}$$

b) At the point a the function $u(x)$ and its derivatives of order $\leq m-1$ vanish. Hence [cf. 2, **15**, formula (23)]

$$u(x)=\frac{1}{(m-1)!}\int_a^x(x-t)^{m-1}u^{(m)}(t)\mathrm{d}t. \tag{32}$$

Differentiating this k times, $k=1, 2, \ldots, s$, we obtain

$$u^k(x)=\frac{1}{(m-k-1)!}\int_a^x (x-t)^{m-k-1}u^{(m)}(t)\,dt. \tag{33}$$

Integrals (32) and (33) fit the general form (6). It suffices to put

$$K(x, t)=\begin{cases}\dfrac{(x-t)^{m-k-1}}{(m-k-1)!}, & a\leq t\leq x.\\ 0, & x<t\leq b.\end{cases}$$

By applying theorem 1 to integral (32) we can satisfy ourselves that it is possible to select from set M a sequence $u_n(x)$, $n=1, 2, \ldots$ which converges in the mean. We substitute $u_n(x)$ for $u(x)$ in (33) and then put $k=1$. By virtue of theorem 1 it is possible to select from the sequence $u_n(x)$ a partial sequence $u_{n_1}, (x), n=1, 2, \ldots$ such that not only this itself but also its first derivatives converge in the mean. We now substitute $u_{n_1}(x)$ for $u(x)$ in (33) and put $k=2$, and again apply theorem 1. Continuing this process, we come to the conclusion that it is possible to select from the given set M a sequence which converges in the mean together with its derivatives of order up to s inclusive. As has been noted previously, there thus follows the discrete nature of the spectrum of equation (27) under boundary conditions (28).

§ 36. The stability of a compressed rod

We consider a rod (in general, of variable section) which is compressed by two longitudinal forces of magnitude P, applied to its ends. The differential equation describing the distorted axis of the rod has the form:

$$E\frac{d^2}{dx^2}\left(I(x)\frac{d^2y}{dx^2}\right)+P\frac{d^2y}{dx^2}=0, \tag{1}$$

where E is Young's Modulus for the material of the rod and $I(x)$ is the geometrical moment of inertia of the section with

abscissa x. With this equation are associated the boundary conditions defining the nature of the fixing at the ends of the rod.

We shall restrict ourselves here to two types of fixing:

1) Both ends of the rod are rigidly fixed. Assuming that the axis of the rod in the unbent state occupies the segment $[0, l]$ of the x-axis, we have in this case

$$y(0)=y(l)=0, \quad y'(0)=y'(l)=0; \tag{2}$$

2) Both ends of the rod are firmly hinged. In this case

$$y(0)=y(l)=0, \quad y''(0)=y''(l)=0. \tag{3}$$

The investigation of other types of fixing does not present any essential difficulties.

The problem of the stability of a compressed rod consists of finding the values of the "critical loads" under which equation (1) has non-trivial solutions which satisfy boundary conditions (2) or (3), respectively. The lowest critical load is of special interest.

Equation (1) is of type (27) § 35, P playing the part of the parameter λ. In this case

$$Ay=E\frac{d^2}{dx^2}\left(I(x)\frac{d^2y}{dx^2}\right), \quad By=-\frac{d^2y}{dx^2}.$$

It is easy to prove by integration by parts that both under conditions (2) and under conditions (3)

$$|y|_A^2=(Ay, y)=E\int_0^l I(x)\left(\frac{d^2y}{dx^2}\right)^2 dx,$$

$$|y|_B^2=(By, y)=\int_0^l\left(\frac{dy}{dx}\right)^2 dx.$$

The problem of the stability of a compressed rod reduces to the problem of seeking the eigenvalues (especially the lowest eigenvalue) of equation (1) under boundary conditions (2) or

(3). Boundary conditions (2) were virtually worked out in the previous section. If no rod section degenerates into a point or a line, so that $I(x)$ nowhere vanishes, then there exists an infinite set of critical loads. The lowest of them is determined by the formula

$$P_1 = \min \frac{\|y\|_A^2}{\|y\|_B^2} = E \cdot \min \frac{\int_0^l I(x) y''^2(x)\, dx}{\int_0^l y'^2(x) dx}; \tag{4}$$

the minimum being sought in the class of functions which satisfy the boundary condition (2).

The case of firm hinging of the ends can also be investigated on the basis of theorems 1 § 34 and 4 § 32; the lowest critical load is determined as before by formula (4) but this time the minimum can be sought in the wider class of functions satisfying only the conditions $y(0)=y(l)=0$; the conditions $y''(0)=$ $=y''(l)=0$ are natural. In general the widening of the class of functions decreases the minimum; hence it follows that, for firm hinging of the ends of the rod, the lowest critical load is less than under rigid fixing.

We can easily reduce the problem of the stability of a rod with hinged ends to finding the eigenvalues of a certain second-order differential equation. We put $Iy''=u$. The function $u(x)$ satisfies the boundary conditions

$$u(0)=u(l)=0 \tag{5}$$

and the differential equation

$$E \frac{d^2u}{dx^2} + \frac{P}{I} u = 0. \tag{6}$$

Equation (6) is of type (24) § 35 for $p(x)=E$, $r(x)=0$, $\varrho(x)=\frac{1}{I}$. The lowest eigenvalue of this equation according to theorem 5 § 34 is the minimum of the ratio

$$E\frac{\int_0^l u'^2 dx}{\int_0^l \frac{u^2}{I} dx}$$

or, what is the same thing, the minimum of the functional

$$E\int_0^l u'^2 dx$$

under the supplementary condition

$$\int_0^l \frac{u^2}{I} dx = 1;$$

the minimum being sought in the class of functions satisfying boundary conditions (5).

We note that if the cross-section of the rod is constant, equation (6) transforms into the well-known Euler equation.

§ 37. Eigenvalues of elliptic operators

The difficulty of the transition from ordinary to partial differential equations in the eigenvalue problem is well known. We cannot successfully carry the investigation further by more or less simple means in all cases.

We shall restrict ourselves to the consideration of non-degenerate elliptic second-order equations and the equation for the bending of a plate.

1. As in the previous section, the basic tool for the investigation here will be theorem 3 § 30. In its turn the earlier application of theorem 1 § 35 also rested upon it. However for the purposes of the present section the latter theorem is insufficient; in its stead we shall use another theorem which we now formulate. We consider a finite domain Ω in m-dimensional space and let $P, Q,$ be points in domain Ω and let r be

the distance between these points. If the constant α lies within the range $0<\alpha<m$ and the function $A(P, Q)$ is bounded, then we shall call the operator

$$Kv=\int_{\Omega}\frac{A(P, Q)}{r^{\alpha}}v(Q)\mathrm{d}\Omega_Q$$

an integral operator with a non-essential singularity, and its kernel $\dfrac{A(P, Q)}{r^{\alpha}}$, *a kernel with a non-essential singularity.*

We then have

THEOREM 1. *An integral operator with a non-essential singularity transforms any set of functions with bounded norms into a set of functions which is compact in the sense of convergence in the mean.* (The proof of this theorem can be found e.g. in the author's book [11].)

2. To be explicit we shall consider that Ω is a domain in three-dimensional space. Let $P(x, y, z)$ and $Q(\xi, \eta, \zeta)$ be points in domain Ω. We consider some function $u(Q)$ which is continuously differentiable in the closed domain $\bar{\Omega}=\Omega+S$ (where, as usual, S denotes the boundary of the domain) and equals zero on S. We surround the point P by a sphere S_ε with centre P and with a radius ε sufficiently small for the whole of this sphere to lie within Ω. We shall denote the remainder of the domain Ω by Ω_ε. We construct the integral

$$\int_{\Omega_\varepsilon}\left\{\nabla_Q\left(\frac{1}{r}\right)\right\}\cdot\nabla u\,\mathrm{d}\Omega_Q=\int_{\Omega_\varepsilon}\left(\frac{\partial u}{\partial\xi}\frac{\partial\dfrac{1}{r}}{\partial\xi}+\frac{\partial u}{\partial\eta}\frac{\partial\dfrac{1}{r}}{\partial\eta}+\frac{\partial u}{\partial\zeta}\frac{\partial\dfrac{1}{r}}{\partial\zeta}\right)\mathrm{d}\Omega_Q$$

and apply formula (8) § 2 to it. It should be remembered here that the boundary of the domain Ω_ε consists of the two surfaces S and S_ε, and therefore our problem will contain two surface integrals. We obtain

$$\int_{\Omega_\varepsilon}\left(\frac{\partial u}{\partial\xi}\frac{\partial\dfrac{1}{r}}{\partial\xi}+\frac{\partial u}{\partial\eta}\frac{\partial\dfrac{1}{r}}{\partial\eta}+\frac{\partial u}{\partial\zeta}\frac{\partial\dfrac{1}{r}}{\partial\zeta}\right)\mathrm{d}\Omega_Q=$$

$$= -\int_{\Omega_\varepsilon} u\nabla^2\frac{1}{r}\,d\Omega_Q + \int_S u\frac{\partial\frac{1}{r}}{\partial n}\,dS + \int_{S_\varepsilon} u\frac{\partial\frac{1}{r}}{\partial n}\,dS_\varepsilon.$$

It is well known (and is easily verified directly) that $\nabla^2\frac{1}{r}=0$. Moreover, by condition, $u|_s=0$, and so the first two integrals on the right vanish and we obtain

$$\int_{S_\varepsilon} u\frac{\partial\frac{1}{r}}{\partial n}\,dS_\varepsilon = \int_{\Omega_\varepsilon}\left(\frac{\partial u}{\partial\xi}\frac{\partial\frac{1}{r}}{\partial\xi}+\frac{\partial u}{\partial\eta}\frac{\partial\frac{1}{r}}{\partial\eta}+\frac{\partial u}{\partial\zeta}\frac{\partial\frac{1}{r}}{\partial\zeta}\right)d\Omega.$$

In formula (8) § 2 which we have used here, n denotes the normal which is external with respect to the domain and so the normal to the sphere S_ε is directed towards its centre and

$$\frac{\partial\frac{1}{r}}{\partial n} = -\left.\frac{\partial\frac{1}{r}}{\partial r}\right|_{r=\varepsilon} = \frac{1}{\varepsilon^2}.$$

We now introduce spherical coordinates θ and φ on the sphere S_ε, $0\leq\theta\leq\pi$, $0\leq\varphi\leq 2\pi$. Then $dS_\varepsilon=\varepsilon^2\sin\theta\,d\theta\,d\varphi$ and $\xi=x+\varepsilon\sin\theta\cos\varphi$, $\eta=y+\varepsilon\sin\theta\sin\varphi$, $\zeta=z+\varepsilon\cos\theta$. Substituting this in the latter equation we reduce it to the form

$$\int_0^{2\pi} d\varphi\int_0^{\pi} u\,(x+\varepsilon\sin\theta\cos\varphi,\ y+\varepsilon\sin\theta\sin\varphi,\ z+\varepsilon\cos\theta)\,d\theta$$

$$= \int_\Omega\left(\frac{\partial u}{\partial\xi}\frac{\partial\frac{1}{r}}{\partial\xi}+\frac{\partial u}{\partial\eta}\frac{\partial\frac{1}{r}}{\partial\eta}+\frac{\partial u}{\partial\zeta}\frac{\partial\frac{1}{r}}{\partial\zeta}\right)d\Omega.$$

We now proceed to the limit by letting $\varepsilon\to 0$. On the right-hand side of the equation this simply leads to making Ω instead of Ω_ε the domain of integration. The left-hand side as a result of passing to the limit assumes the form

$$\int_0^{2\pi} d\varphi\int_0^{\pi} u(x,y,z)\sin\theta\,d\theta = u(x,y,z)\int_0^{2\pi} d\varphi\int_0^{\pi}\sin\theta\,d\theta$$

$$= 4\pi u(x,y,z) = 4\pi u(P).$$

Dividing both sides of the equation by 4π, we obtain the important formula

$$u(P)=\frac{1}{4\pi}\int_{\Omega}\left(\frac{\partial u}{\partial \xi}\frac{\partial \frac{1}{r}}{\partial \xi}+\frac{\partial u}{\partial \eta}\frac{\partial \frac{1}{r}}{\partial \eta}+\frac{\partial u}{\partial \zeta}\frac{\partial \frac{1}{r}}{\partial \zeta}\right)d\Omega, \qquad (1)$$

which is valid if the function u vanishes on the boundary of the domain Ω.

The integral in (1) can be divided in a natural fashion into three parts, it is easily seen that the kernel of each of them has a non-essential singularity. Thus, for example,

$$\frac{\partial \frac{1}{r}}{\partial \xi}=\frac{x-\xi}{r^3}=\frac{x-\xi}{r}\frac{1}{r^2};$$

where the function

$$A(P,Q)=\frac{x-\xi}{r}$$

is bounded (since $|x-\xi|\leq r$ and thus $|A(P, Q)|\leq 1$) and the exponent $\alpha=2$ is less than the number of variables of the space.

It is also possible to construct a formula analogous to formula (1) when the number of variables of the space does not equal 3; in particular if Ω is a surface domain and the function $u(x, y)$ vanishes on its boundary, then

$$u(x, y)=\frac{1}{2\pi}\int_{\Omega}\left(\frac{\partial u}{\partial \xi}\frac{\partial \ln\frac{1}{r}}{\partial \xi}+\frac{\partial u}{\partial \eta}\frac{\partial \ln\frac{1}{r}}{\partial \eta}\right)d\Omega. \qquad (1_1)$$

3. We now consider the problem of the eigenvalues of a differential operator of elliptic type

$$Lu=-\sum_{j,\,k=1}^{m}\frac{\partial}{\partial x_j}\left(A_{jk}\frac{\partial u}{\partial x_k}\right)+Cu \qquad (2)$$

under the condition that, on the boundary of the domain Ω considered, the condition $u=0$ is satisfied. We retain the stipulations of § 24 concerning the coefficients A_{jk} and C, so that the operator L is positive-bounded-below. In the particular case where

$$A_{jj}=1; \quad A_{jk}=0, \ j\neq k,$$

we obtain the problem of eigenvalues for the Laplacian.

The proof, that the condition of theorem 3 § 31 is satisfied for this operator was given by the author in [12]. By formulae (7) and (9) § 24, we have

$$|u|^2=(Lu, u)\geq\mu_0\int\limits_{\Omega}\sum_{j=1}^{m}\left(\frac{\partial u}{\partial\xi_j}\right)^2 d\Omega.$$

Putting the number of spatial variables $m=3$, for example, we can rewrite the latter equation

$$|u|^2\geq\mu_0\int\limits_{\Omega}\left[\left(\frac{\partial u}{\partial\xi}\right)^2+\left(\frac{\partial u}{\partial\eta}\right)^2+\left(\frac{\partial u}{\partial\zeta}\right)^2\right]d\Omega. \tag{3}$$

Let there be given a set M of functions each of whose energy norms is bounded $|u|\leq C$, $u\in M$. From inequality (3) it then follows that

$$\left\|\frac{\partial u}{\partial\xi}\right\|=\left\{\int\limits_{\Omega}\left(\frac{\partial u}{\partial\xi}\right)^2 d\Omega\right\}^{\frac{1}{2}}\leq\frac{C}{\sqrt{(\mu_0)}};$$

$$\left\|\frac{\partial u}{\partial\eta}\right\|\leq\frac{C}{\sqrt{(\mu_0)}}, \ \left\|\frac{\partial u}{\partial\zeta}\right\|\leq\frac{C}{\sqrt{(\mu_0)}}. \tag{4}$$

Since $\frac{\partial\frac{1}{r}}{\partial\xi}$ is a kernel with a non-essential singularity and the norms of the functions $\frac{\partial u}{\partial\xi}$ are all bounded it follows that, if u is any function from the set M, it is possible by virtue of

theorem 1, to select from set M a sequence $u_{1n}(P)$, $n=1, 2, \ldots$ such that the integral

$$\frac{1}{4\pi}\int_{\Omega}\frac{\partial u_{1n}}{\partial \xi}\frac{\partial \frac{1}{r}}{\partial \xi}\,d\Omega \tag{5}$$

converges in the mean to some limit. The norms of the functions $\frac{\partial u_{1n}}{\partial \eta}$ are all bounded since $u_{1n}\in M$ and thus, by virtue of the inequality (4)

$$\left\| \frac{\partial u_{1n}}{\partial \eta} \right\| \leq \frac{C}{\sqrt{(\mu_0)}}.$$

Again applying theorem 1, we find that it is possible to select from the sequence $u_{1n}(P)$ the partial sequence $u_{2n}(P)$ such that the integral

$$\frac{1}{4\pi}\int_{\Omega}\frac{\partial u_{2n}}{\partial \eta}\frac{\partial \frac{1}{r}}{\partial \eta}\,d\Omega$$

converges in the mean. In this case, obviously, the sequence

$$\frac{1}{4\pi}\int_{\Omega}\frac{\partial u_{2n}}{\partial \xi}\frac{\partial \frac{1}{r}}{\partial \xi}\,d\Omega,$$

which constitutes part of sequence (5) also converges in the mean – its limit being the limit of sequence (5). Similarly we find that it is possible to select from the sequence u_{2n} a new partial sequence which we denote simply by $u_n(P)$, $n=1, 2, \ldots$ such that all three integrals

$$\frac{1}{4\pi}\int_{\Omega}\frac{\partial u_n}{\partial \xi}\frac{\partial \frac{1}{r}}{\partial \xi}\,d\Omega,\ \frac{1}{4\pi}\int_{\Omega}\frac{\partial u_n}{\partial \eta}\frac{\partial \frac{1}{r}}{\partial \eta}\,d\Omega,\ \frac{1}{4\pi}\int_{\Omega}\frac{\partial u_n}{\partial \zeta}\frac{\partial \frac{1}{r}}{\partial \zeta}\,d\Omega \tag{6}$$

converge to certain limits as $n\to\infty$.

But from formula (1) it is clear that the functions $u_n(P)$ themselves also converge in the mean to a limit which equals the sum of the limits of integrals (6). Thus, we have succeeded in selecting from the set M of functions, which are bounded in energy, a sequence of functions $u_n(P)$ which converges in the mean. On the basis of theorem 3 § 31 we now conclude that *under the boundary condition* $u|_s=0$ *the non-degenerate elliptic operator* (2) *has a discrete spectrum in the finite domain* Ω, *the lowest eigenvalue of this operator being defined by the formula*

$$\lambda_1=\min\int_{\Omega}\left(\sum_{j,k=1}^{m}A_{jk}\frac{\partial u}{\partial x_j}\frac{\partial u}{\partial x_k}+Cu^2\right)d\Omega,$$

where the minimum is sought in the class of functions which vanish on the boundary of the domain and satisfy the equation

$$\int_{\Omega}u^2d\Omega=1.$$

In particular, the lowest eigenvalue of the operator $-\nabla^2$ (∇^2 is the Laplacian) equals the minimum of the integral

$$\int_{\Omega}\sum_{i=1}^{m}\left(\frac{\partial u}{\partial x_i}\right)^2d\Omega$$

with respect to the class of functions just mentioned.

4. The problem of the eigenfrequencies of the flexural oscillations of a plate with rigid fixing or free support reduces to the problem of the eigenvalues of the biharmonic operator ∇^4. The corresponding eigenfunctions satisfy the boundary conditions

$$w|_S=0,\quad \left.\frac{\partial w}{\partial n}\right|_S=0, \tag{7}$$

if the edge of the plate is rigidly fixed, and the conditions

$$w_S|=0,\quad \left[\nabla^2w-\frac{1-\sigma}{\varrho}\frac{\partial w}{\partial n}\right]_S=0, \tag{8}$$

if the edge is freely supported. From formulae (16), (17) and (25) § 27 it follows that under both conditions of fixing, there occurs the inequality

$$|w|^2 \geq c \int_S \int \left[\left(\frac{\partial w}{\partial x} \right)^2 + \left(\frac{\partial w}{\partial y} \right)^2 \right] dS, \tag{9}$$

where c is a positive constant, equal to $\varkappa$ for the conditions (7) and $\frac{1-\sigma}{A}$ for the conditions (8). As in § 27, we denote by S the domain occupied by the plate.

Let there be given a set M of functions each of whose energy norms is bounded $|w| \leq C$, $w \in M$. By inequality (9)

$$\left\| \frac{\partial w}{\partial x} \right\| \leq \frac{C}{\sqrt{(c)}}, \qquad \left\| \frac{\partial w}{\partial y} \right\| \leq \frac{C}{\sqrt{(c)}}.$$

It now follows from formula (1_1) and theorem 1 that it is possible to select from set M a sequence which converges in the mean; in other words the set M is compact in the sense of convergence in the mean.

From theorem 3 § 31 it then follows that *both in the case of rigid fixing of the edge and in the case of free support of the edge the biharmonic operator has a discrete spectrum.*

5. We now make the following observation. The lowest eigenvalue of the operator ∇^4 under the conditions (7) will be denoted by λ_1, and that under conditions (8) by μ_1. By the general formula (3) § 33 it is possible to determine both numbers λ_1 and μ_1 as the minimum of the ratio

$$\frac{|w|^2}{\|w\|^2},$$

with the difference that in the first case the minimum is sought with respect to functions which satisfy conditions (7) and in the second with respect to functions which satisfy only the first of conditions (8), the second condition (8) being natural.

In the case of a freely supported edge [formula (21) § 28]

$$|w|^2=\int_S\int\left\{\left(\frac{\partial^2 w}{\partial x^2}\right)^2+2\sigma\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}+\left(\frac{\partial^2 w}{\partial y^2}\right)^2 +2(1-\sigma)\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2\right\}dS, \tag{10}$$

and in the case of a rigidly fixed edge [formula (16) § 27]

$$|w|^2=\int_S\int\left\{\left(\frac{\partial^2 w^2}{\partial x^2}\right)^2+2\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2+\left(\frac{\partial^2 w}{\partial y^2}\right)^2\right\}dS. \tag{11}$$

However by multiplying equation (6_2) § 27 by -2σ and adding it to (11) we can satisfy ourselves that formula (10) is also applicable in the case of rigid fixing of the edge.

From what has been said it follows that λ_1 and μ_1, are the minima of one and the same ratio

$$\frac{\int_S\int\left\{\left(\frac{\partial^2 w}{\partial x^2}\right)^2+2\sigma\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}+\left(\frac{\partial^2 w}{\partial y^2}\right)^2+2(1-\sigma)\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2\right\}dS}{\int_S\int w^2 dS}, \tag{12}$$

but the first minimum is sought for a more restricted class of functions. Hence it follows that $\lambda_1\geq\mu_1$, i.e. that the lowest eigenfrequency for the vibration of a freely supported plate does not exceed the lowest eigenfrequency for the oscillations of a plate which is rigidly fixed at its edge. In § 40 we shall return to the question in a more general form.

6. It can be proved that, in the case of a finite domain Ω, an elliptic non-degenerate operator also has a discrete spectrum under the other boundary conditions considered in § 24, but the proof here is considerably more complex. It is given, for example, in the author's book [11].

7. One can readily cite examples of differential operators which do not possess a discrete spectrum. Without giving the proof, we note that in the case of an infinite domain Ω the

spectrum of the Laplacian operator is not discrete. What is more, in this case the Laplacian has no eigenfunctions with a finite norm. Under certain conditions a degenerate elliptic operator does not have a discrete spectrum even in the case where Ω is a finite domain (see the author's paper [16]).

§ 38. The stability of a compressed plate

We shall consider the bending of a plate in the plane of which the stresses T_x, T_{xy}, T_y act. We shall consider these stresses to be compressional, understanding by this that, at any point of the plate, the principle normal stresses T_1 and T_2 are compressional: $T_1 < 0$, $T_2 < 0$. We shall also suppose that both T_1 and T_2 are always greater in absolute value than some positive constant T_0 so that $|T_1| \geq T_0$, $|T_2| \geq T_0$, or, as T_1 and T_2 are negative,

$$T_1 \leq -T_0, \quad T_2 \leq -T_0.$$

We noted in § 28 that the expression

$$T_x\left(\frac{\partial w}{\partial x}\right)^2 + 2T_{xy}\frac{\partial w}{\partial x}\frac{\partial w}{\partial y} + T_y\left(\frac{\partial w}{\partial y}\right)^2$$

is invariant under a rotation of the coordinate axes. By rotating the coordinate axes so that they coincide with the principle axes of stress, we find

$$T_x\left(\frac{\partial w}{\partial x}\right)^2 + 2T_{xy}\frac{\partial w}{\partial x}\frac{\partial w}{\partial y} + T_y\left(\frac{\partial w}{\partial y}\right)^2 = T_1\left(\frac{\partial w}{\partial x_1}\right)^2 + T_2\left(\frac{\partial w}{\partial x_2}\right)^2$$

$$\leq -T_0\left[\left(\frac{\partial w}{\partial x_1}\right)^2 + \left(\frac{\partial w}{\partial x_2}\right)^2\right].$$

The expression

$$\left(\frac{\partial w}{\partial x_1}\right)^2 + \left(\frac{\partial w}{\partial x_2}\right)^2 = (\nabla w)^2$$

is also invariant under rotation of the axes and so

$$\left(\frac{\partial w}{\partial x_1}\right)^2 + \left(\frac{\partial w}{\partial x_2}\right)^2 = \left(\frac{\partial w}{\partial x}\right)^2 + \left(\frac{\partial w}{\partial y}\right)^2,$$

and finally

$$T_x\left(\frac{\partial w}{\partial x}\right)^2+2T_{xy}\frac{\partial w}{\partial x}\frac{\partial w}{\partial y}+T_y\left(\frac{\partial w}{\partial y}\right)^2\leq -T_0\left[\left(\frac{\partial w}{\partial x}\right)^2+\left(\frac{\partial w}{\partial y}\right)^2\right]. \quad (1)$$

We shall assume that the edge of the disc is rigidly fixed:

$$w|_L=0, \qquad \left.\frac{\partial w}{\partial n}\right|_L=0. \quad (2)$$

Inequality (1) together with formulae (4) and (9) § 28 show that the operator $-T$ is positive-bounded-below.

We now suppose that the stresses, while remaining compressional, are proportional to some parameter λ, which consequently must be positive. As was explained in § 29 the plate can lose stability at those values of λ which are eigenvalues of equation (12) § 29. If we introduce the notation

$$\tilde{T}\omega=-\frac{h}{D}T\omega, \quad (3)$$

then this equation can be written in the form

$$\nabla^4\omega-\lambda\tilde{T}\omega=0. \quad (4)$$

We note that the operator $\tilde{T}$ is positive-bounded-below.

We now prove that equation (4) under the conditions (2) has an infinite set of positive eigenvalues. It follows therefrom that there exists an infinite set of values of the parameter at which a plate with a rigidly fixed edge loses stability.

Equation (4) fits the general type of equation of § 34, if we put $A=\nabla^4$, $B=\tilde{T}$. We shall have proved our assertion, if, in correspondence with theorem 4 § 34, we prove that any set bounded in the energy of operator A is compact in the sense of convergence in the energy of operator B.

In § 27 it was established [formula (16)] that

$$\|w\|_A^2=\int_S\int\left\{\left(\frac{\partial^2 w}{\partial x^2}\right)^2+2\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2+\left(\frac{\partial^2 w}{\partial y^2}\right)^2\right\}dS. \quad (5)$$

Furthermore,

$$\|w\|_B^2=(\tilde{T}w,\ w)=-\frac{h}{D}(Tw,\ w).$$

By formula (10) § 28 we now have

$$\|w\|_B^2 \leq \frac{2Nh}{D} \int_S \int \left\{ \left(\frac{\partial w}{\partial x} \right)^2 + \left(\frac{\partial w}{\partial y} \right)^2 \right\} dS, \tag{6}$$

where N denotes the largest of the maxima of the quantities $|T_x|$, $|T_{xy}|$, $|T_y|$.

Now let there be given a set M of functions which satisfy conditions (2) and let $\|w\| \leq c$, if $w \in M$. From formula (5) we then readily conclude that

$$\int_S \int \left\{ \left(\frac{\partial^2 w}{\partial x^2} \right)^2 + \left(\frac{\partial^2 w}{\partial x \partial y} \right)^2 \right\} dS \leq C^2, \tag{7}$$

$$\int_S \int \left\{ \left(\frac{\partial^2 w}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 w}{\partial y^2} \right)^2 \right\} dS \leq C^2. \tag{8}$$

From conditions (2) it follows that

$$\left. \frac{\partial w}{\partial x} \right|_L = 0, \qquad \left. \frac{\partial w}{\partial y} \right|_L = 0. \tag{9}$$

As usual, let ∇^2 denote the Laplacian. We consider the positive-bounded-below operator $-\nabla^2$, specified for functions which vanish on the edge L.

From (9) it is clear that if $w \in M$, $\dfrac{\partial w}{\partial x}$ and $\dfrac{\partial w}{\partial y}$ enter into the field of definition of the operator $-\nabla^2$ and inequalities (7) and (8) show that these derivatives are all bounded in the energy of this operator. From the results of § 37 it then follows that the sets of derivatives $\dfrac{\partial w}{\partial x}$ and $\dfrac{\partial w}{\partial y}$ are compact in the sense of convergence in the mean. Thus, it is possible to select from the set M a sequence w_n, $n = 1, 2, \ldots$ such that the corresponding sequences $\dfrac{\partial w_n}{\partial x}$ and $\dfrac{\partial w_n}{\partial y}$ converge in the mean. But then

$$\lim_{n, m \to \infty} \int_S \int \left\{ \left(\frac{\partial w_n}{\partial x} - \frac{\partial w_m}{\partial x} \right)^2 dx dy = 0, \right.$$

$$\lim_{n,\,m\to\infty}\int_S\int\left(\frac{\partial w_n}{\partial y}-\frac{\partial w_m}{\partial y}\right)^2 dx\,dy=0.$$

Now from the inequality (6) it follows that

$$\|w_n-w_m\|_B^2\leq\frac{2Nh}{D}\int_S\int\left\{\left(\frac{\partial w_n}{\partial x}-\frac{\partial w_m}{\partial x}\right)^2+\left(\frac{\partial w_n}{\partial y}-\frac{\partial w_m}{\partial y}\right)^2\right\}dS\xrightarrow[n,\,m\to\infty]{}0,$$

and this means that the sequence w_n, $n=1, 2, \ldots$ converges in the energy of the operator B, and our assertion is proved.

The lowest critical value of the parameter can be found as the minimum of integral (5) under the condition

$$\|w\|_B^2=-\frac{h}{D}\int_S\int\left\{T_x\left(\frac{\partial w}{\partial x}\right)^2+2T_{xy}\frac{\partial w}{\partial x}\frac{\partial w}{\partial y}+T_y\left(\frac{\partial w}{\partial y}\right)^2\right\}dS=1. \quad (10)$$

The minimum is sought with respect to the set of functions which satisfy conditions (2) for the rigid fixing of the edge. Using S. L. Sobolev's embedding theorem, it is also possible to prove the existence of an infinite set of critical values of the parameter for the other commonly encountered edge fixing conditions.

§ 39. The characteristic vibrations of an elastic body

In the present section we shall consider the problem of the characteristic vibrations of an elastic body with a fixed boundary, and we shall prove that such a body has an infinite set of eigenfrequencies. In order to simplify the formulation we abandon the notation of § 26 and use the following notation: S is the boundary of the domain Ω; $\bar{\Omega}=\Omega+S_1$, P, Q are points with the Cartesian coordinates x_1, x_2, x_3 and ξ_1, ξ_2, ξ_3 respectively. Further $\mathbf{u}$ is the displacement vector; we denote

the components of this vector along the axes x_1, x_2, x_3 of the Cartesian coordinates by u_1, u_2, u_3. We further put

$$s_{ik}=s_{ki}=\frac{1}{2}\left(\frac{\partial u_i}{\partial x_k}+\frac{\partial u_k}{\partial x_i}\right). \tag{1}$$

These quantities are equal to the components of the deformation tensor divided by 2. We shall denote the components of the stress tensor by τ_{ik}; They satisfy the equation which expresses the generalized Hooke's law

$$\tau_{ik}=\sum_{l,\,m=1}^{3} c_{iklm}s_{lm}=\tau_{ki}; \tag{2}$$

the elasticity coefficients c_{iklm} being subject to the well-known symmetry relations

$$c_{iklm}=c_{kilm}=c_{ikml}=c_{lmik},$$

so that the number of distinct coefficients c_{iklm} is not greater than 21. The stress components satisfy the well-known equations of motion

$$\gamma\frac{\partial^2 u_i}{\partial t^2}-\sum_{k=1}^{3}\frac{\partial \tau_{ik}}{\partial x_k}-K_i=0, \quad i=1,\ 2,\ 3,$$

where t is the time, γ is the density of the elastic body, the K_i are the components of the vector of the force per unit volume $\mathbf{K}$. Substituting in this the values of τ_{ik} from (2), we obtain the equations which the elastic displacement vector satisfies. We write these equations in the form of a single vector equation:

$$\gamma\frac{\partial^2 \mathbf{u}}{\partial t^2}+\mathbf{Au}-\mathbf{K}=0, \tag{3}$$

where

$$\mathbf{Au}=-\sum_{i,k,l,\,m=1}^{3}\frac{\partial}{\partial x_k}(c_{iklm}s_{lm})\mathbf{x}_i^{(0)}; \tag{4}$$

$\mathbf{x}_i^{(0)}$ is unit vector along the x_i-axis. If the medium is homogeneous then

$$\mathbf{Au} = -\sum_{i,k,l,m=1}^{3}{}' c_{iklm}\frac{\partial s_{lm}}{\partial x_k}\mathbf{x}_i^{(0)}.$$

In the equilibrium case $\dfrac{\partial^2 \mathbf{u}}{\partial t^2}=0$ and we arrive at the equation of elastic equilibrium

$$\mathbf{Au}=\mathbf{K}. \tag{5}$$

We obtain the equation for the free vibrations by assuming that $\mathbf{K}\equiv 0$ and by replacing $\mathbf{u}$ by $\mathbf{u}(P)e^{i\omega t}$. We then obtain

$$\mathbf{Au}-\lambda\mathbf{u}=0, \quad \lambda=\gamma\omega^2. \tag{6}$$

We shall call operator $\mathbf{A}$ the elasticity theory operator.

We introduce the symmetric tensor V_k, the components of which are

$$v_{ik}=\frac{1}{12\pi}\left\{\frac{5}{r}\delta_{ik}+\frac{r^2}{2}\frac{\partial^2}{\partial x_i \partial x_k}\left(\frac{1}{r}\right)\right\}, \tag{7}$$

where r is the distance between the points $P(x_1, x_2, x_3)$ and $Q(\xi_1,\xi_2,\xi_3,)$ and $\delta_{ik}=\begin{cases}0, i\neq k,\\ 1, i=k.\end{cases}$

V is a particular case of the well-known Somilyan tensor; (see, e.g. E. Trefftz [2]). Each column $\mathbf{v}_i$ of the tensor V satisfies a homogeneous equation of the theory of an elastic isotropic medium, the Lamé constants of which have the values $\lambda=0$ and $\mu=\dfrac{1}{2}$. We shall denote the corresponding operator of elasticity theory by $\mathbf{A}_0$ so that $\mathbf{v}_i$ satisfies the equation

$$\mathbf{A}_0\mathbf{v}_i=0. \tag{8}$$

We enclose the point P in a sphere of radius ε and we apply Betti's first formula to the remainder Ω_ε of the domain Ω by putting $\mathbf{u}'=\mathbf{u}$, $\mathbf{u}''=\mathbf{v}$, $\mathbf{A}=\mathbf{A}_0$ where $\mathbf{u}$ is any vector which vanishes on the boundary of the domain Ω. In this case the integral over S disappears. In fact the integral on the left in

Betti's formula also vanishes by virtue of equation (8), and we arrive at the equation

$$\int_{\Omega_\varepsilon} w(\mathbf{u}, \mathbf{v}_i)\mathrm{d}\Omega + \int_{S_\varepsilon} \mathbf{t}_{i,\varepsilon}\mathbf{u}\mathrm{d}S = 0, \tag{9}$$

where S_ε is the surface of the sphere of radius ε and with centre P mentioned above, and $\mathbf{t}_{i,\varepsilon}$ is the stress vector over the surface of the sphere S_ε corresponding to the displacement $\mathbf{v}_i$. Letting $\varepsilon \to 0$ in (9), we obtain, as in § 37,

$$u_i(P) = \int_\Omega W(\mathbf{u}, \mathbf{v}_i)\mathrm{d}\Omega.$$

It is easily seen that for $\lambda = 0$, $\mu = \dfrac{1}{2}$

$$W(\mathbf{u}, \mathbf{v}_i) = \sum_{l,m=1}^{3} s_{lm}s_{lm}^{(i)},$$

where $2s_{lm}$ and $2s_{lm}^{(i)}$ are the components of deformation corresponding to the displacements $\mathbf{u}$ and $\mathbf{v}_i$. Thus

$$u_i(P) = \int_\Omega \sum_{l,m=1}^{3} s_{lm}s_{lm}^{(i)}\mathrm{d}\Omega. \tag{10}$$

Integral (10) is the sum of integrals of the form

$$\psi_{lm}^{(i)}(P) = \int_\Omega s_{lm}s_{lm}^{(i)}\mathrm{d}\Omega \tag{11}$$

the kernels of which obey the relation

$$\left|s_{lm}^{(i)}\right| \leq \frac{C}{r^2}, \qquad C = \text{const}$$

and belong, consequently, to the class of kernels with a non-essential singularity. But then, by theorem 1 § 27, the integral operator (11) transforms any set of functions s_{lm} with bounded norm into a set of functions $\psi_{lm}^{(i)}$ which is compact in the sense of convergence in the mean.

Starting from the fact that the kernel of integral (11) has a non-essential singularity, it is possible to establish the inequality $\| \varphi_{kl}^{(i)} \| \leq C_1 \| s_{kl} \|$, $C_1 = \text{const}$. Adding, we obtain

$$\|\mathbf{u}\|^2 \leq C^2 \sum_{i,\,k,\,l=1}^{3} \| \varphi_{k,\,l}^{(i)} \|^2 \leq C_3 \sum_{k,\,l=1}^{3} \|s_{kl}\|^2.$$

By using the positive definite character of the quantity $W(\mathbf{u})$ we can prove that, using the theory of quadratic forms (see e.g. I. M. Gel'fand [1]),

$$\sum_{k,\,l=1}^{3} s_{kl}^2 \leq C_4 W(\mathbf{u}).$$

But then

$$\|\mathbf{u}\|^2 \leq 2C_5 \int_{\Omega} W(\mathbf{u}) \mathrm{d}\Omega, \quad 2C_5 = C_3 C_4,$$

or, by formula (14) § 26, $\|u\|^2 \leq C_5 (\mathbf{Au}, \mathbf{u})$, whence

$$(\mathbf{Au}, \mathbf{u}) \geq \frac{1}{C_5} \|\mathbf{u}\|^2, \tag{12}$$

i.e. *in the case of a fixed boundary for the body, the elasticity theory operator is positive-bounded-below.*

It has already been noted that the operator (11) transforms a bounded set of functions into a compact one. From the formula (10) it then follows that the set of displacement vectors, which vanish on the boundary of the body, will be compact if the corresponding integrals of the energy are all bounded. There now follows from theorem 3 § 31 the discrete nature of the spectrum of the elasticity theory operator in the case of rigid fixing of the boundary of the body.

The elasticity theory operator also has a discrete spectrum under other commonly stipulated boundary conditions. A detailed discussion of this is given in the author's book [11]. We restrict ourselves here to giving a few bibliographical references. For an arbitrary piecewise smooth boundary the discrete nature of the spectrum of the elasticity theory operator was established by K. Friedrichs [3] in the case of a rigidly fixed or a free boundary, this result having been obtained far earlier by H. Weyl on the assumption of a sufficiently smooth boundary [1]. D. M. Eidus [1, 2] gave another proof,

considerably simpler than Friedrichs's, for the case of a free boundary and extended it to a few other boundary value problems in elasticity theory. The proof presented in the text belongs to the author [12].

In conclusion we note that the lowest eigenvalue of the elasticity theory operator under the condition of a rigid fixing of the boundary of the body equals the minimum of twice the potential energy of deformation

$$2\int_{\Omega} W(\mathbf{u})\, d\Omega \tag{13}$$

under the condition that

$$\int_{\Omega} \mathbf{u}^2 d\Omega = 1. \tag{14}$$

The minimum is sought with respect to displacement vectors which vanish on the boundary of the domain. We further state, without presenting the proof, that the lowest eigenvalue of the elasticity theory operator in the case of a free boundary also equals the minimum of integral (13) but the displacement vectors in this case satisfy, in addition to equation (14), the further equation ($\mathbf{R}$ is the radius vector of the point P)

$$\int_{\Omega} \mathbf{u} d\Omega = 0, \quad \int_{\Omega} \mathbf{R} \times \mathbf{u} d\Omega = 0;$$

there being no need to subject these vectors to any boundary conditions, since the conditions at a free boundary are natural.

§ 40. The minimax principle and its implications

1. Let A be a symmetric positive definite operator with a discrete spectrum and let $\lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \ldots$ be its eigenvalues and $\varphi_1(P), \varphi_2(P), \varphi_3(P), \ldots$ the corresponding eigenfunctions. By the definition of a discrete spectrum, the system of eigenfunctions is complete in the sense of convergence in the mean,

and so if $u(P)$ is any function with a finite norm, then it is possible to expand it in an orthogonal series

$$u(P)=\sum_{n=1}^{\infty}a_n\varphi_n(P), \quad a_n=(u, \varphi_n). \tag{1}$$

We further suppose that $u\in D_A$, and then the function Au has in each case a finite norm and also expands into an orthogonal series,

$$Au=\sum_{n=1}^{\infty}(Au, \varphi_n)\varphi_n(P).$$

The coefficients of this series are readily evaluated. Since A is a symmetric operator, $(Au, \varphi_n)=(u, A\varphi_n)$. In its turn the eigenfunction φ_n satisfies the equation $A\varphi_n=\lambda_n\varphi_n$. Hence

$$(Au, \varphi_n)=(u, \lambda_n\varphi_n)=\lambda_n(u, \varphi_n)=\lambda_n a_n,$$

and, consequently,

$$Au=\sum_{n=1}^{\infty}\lambda_n a_n\varphi_n(P). \tag{2}$$

Series (2) converges in the mean, and so, it is possible to subject it to scalar multiplication by any function with a finite norm (see theorem 2 § 7). Multiplying both sides of equation (2) scalarly by u, we obtain

$$(Au, u)=\sum_{n=1}^{\infty}\lambda_n a_n(\varphi_n, u),$$

or, since $(\varphi_n, u)=(u, \varphi_n)=a_n$,

$$(Au, u)=\sum_{n=1}^{\infty}\lambda_n a_n^2. \tag{3}$$

As we already know, the lowest eigenvalue λ_1, of operator A equals the minimum of the quantity (Au, u) under the condition that $\|u\|=1$. This result is extremely readily obtained anew from formula (3). In fact, by Parseval's equation (formula 18 § 10)

$$\|u\|^2=\sum_{n=1}^{\infty}a_n^2,$$

and the condition $\|u\|=1$ assumes the form

$$\sum_{n=1}^{\infty} a_n^2 = 1. \tag{4}$$

In formula (3) we replace all the numbers λ_n by the lowest of them, λ_1. This gives us

$$(Au, u) \geq \lambda_1 \sum_{n=1}^{\infty} a_n^2, \text{ or } \lambda_1 \leq (Au, u). \tag{5}$$

On the other hand inequality (5) becomes an equality if we put $a_1=1$, $a_2=a_3=\ldots=0$ i. e., if we put $u(P)=\varphi_1(P)$. Thus we have proved that $\lambda_1=\min(Au, u)$ under the condition $\|u\|=1$.

Formula (3) gives a new method of evaluating the successive eigenvalues of operator A. It might be thought that this method is of little use in the evaluation of the eigenvalues λ_n, but its theoretical value is considerable, and it allows us to establish several important properties of eigenvalues.

Let $v_1(P)$, $v_2(P)$, $\ldots$, $v_{k-1}(P)$ be arbitrary functions with finite norm. We pose the problem of finding the minimum of the functional (Au, u) under the conditions

$$\|u\|^2 = \sum_{n=1}^{\infty} a_n^2 = 1 \tag{4$_1$}$$

and

$$(u, v_1)=0; \quad (u, v_2)=0; \quad \ldots; \quad (u, v_{k-1})=0. \tag{6}$$

We denote this minimum by $\lambda(v_1, v_2, \ldots, v_{k-1})$. We shall prove that

$$\lambda(v_1, v_2, \ldots, v_{k-1}) \leq \lambda_k \tag{7}$$

and that the equality sign in (7) is achieved under a certain special choice of the functions $v_1(P)$, $v_2(P)$, $\ldots$, $v_{k-1}(P)$.

We expand $v_1, v_2, \ldots, v_{k-1}$ into series in the eigenfunctions φ_n, and let

$$v_i = \sum_{n=1}^{\infty} b_{in}\varphi_n(P), \quad i=1, 2, \ldots, k-1.$$

Condition (6) can then be represented in the form

$$\sum_{n=1}^{\infty} b_{in}a_n = 0, \quad i=1, 2, \ldots, k-1. \tag{8}$$

We select the function $\bar{u}(P)$ so that $a_{k+1}=a_{k+2}=\ldots=0$. Then condition (8) reduces to a homogeneous system of linear algebraic equations,

$$\sum_{n=1}^{k} b_{in}a_n=0, \quad i=1, 2, \ldots, k-1 \tag{9}$$

in the unknowns $a_1, a_2, \ldots, a_k$ in which the number of equations, $k-1$, is less than the number of unknowns, k. As is well known such a system has an infinite set of solutions from which it is possible to select at least one such that

$$\sum_{n=1}^{k} a_n^2=1,$$

and $\|\bar{u}\|=1$. In this case, by virtue of relation (3):

$$(A\bar{u}, \bar{u})=\sum_{n=1}^{k}\lambda_n a_n^2.$$

Replacing all the λ_n here by the λ_k (which are greater), we obtain

$$(A\bar{u}, \bar{u})\leq\lambda_k\sum_{n=1}^{k}a_n^2=\lambda_k.$$

Thus with the choice of the function $u(P)$ which satisfies (4_1) and (6), namely, with $u(P)=\bar{u}(P)$ we have $(Au, u)\leq\lambda_k$. Also the minimal value of (Au, u) will not be larger than λ_k under the condition mentioned. Thus it follows that

$$\lambda(v_1, v_2, \ldots, v_{k-1})\leq\lambda_k.$$

It is easy to prove that the equality sign is admissible. In fact, it suffices to put

$$v_i(P)=\varphi_i(P), \quad i=1, 2, \ldots, k-1,$$

and the conditions (6) transform into condition (9) § 31 with the aid of which the numbers λ_k are defined.

From what has been said follows a method of determining eigenvalues which is known by the name of *the minimax*

principle: the k-th eigenvalue λ_k *of a symmetric operator* A *with a discrete spectrum equals the maximum of the quantity* $\lambda(v_1, v_2, \ldots, v_{k-1})$ *for all possible systems of functions* $v_1, v_2, \ldots, v_{k-1}$. *In its turn the quantity* $\lambda(v_1, v_2, \ldots, v_{k-1})$ *equals the minimum of the quantity* (Au, u) *under conditions* (4) *and* (6).

NOTE: Formula (3) can be represented in the form

$$\|u\|^2 = \sum_{n=1}^{\infty} \lambda_n a_n^2.$$

It is readily proved (we leave this to the reader) that the quantity $\lambda(v_1, v_2, \ldots, v_{k-1})$ can be obtained as the minimum of the quantity $\|u\|^2$, where u belongs to the set of functions with finite energy which satisfy the normality condition (4) and the orthogonality conditions (6).

2. Let A and B be two positive definite operators.

We shall say that *operator* A *is not less than operator* B, if any function having finite energy with respect to operator A also has finite energy with respect to operator B, and for any such function $\|u\|_A \geq \|u\|_B$. If in addition $A \neq B$ we shall say that $A > B$.

THEOREM 1. *Let* A *and* B *be positive definite operators, and let* $A \geq B$ *and let the spectra of both operators be discrete. Then for any* k *the* k*-th eigenvalue of operator* B *does not exceed the* k*-th eigenvalue of operator* A.

We denote by λ_k and μ_k the eigenvalues of operators A and B respectively. Further we denote by $\lambda(v_1, v_2, \ldots, v_{k-1})$ and $\mu(v_1, v_2, \ldots, v_{k-1})$ the minima of the quantities $\|u\|_A^2$ and $\|u\|_B^2$ under conditions (4) and (6). The inequality $A \geq B$ means that for any function $u(P)$ the inequality $\|u\|_A^2 \geq \|u\|_B^2$ is valid; it is also valid consequently for any function satisfying conditions (4) and (6) mentioned above.

But then, obviously, the minima of the quantities $\|u\|_A^2$ and $\|u\|_B^2$ are also connected by the same inequality, i.e.

$$\lambda(v_1, v_2, \ldots, v_{k-1}) \geq \mu(v_1, v_2, \ldots, v_{k-1}).$$

Hence it also follows that

$$\max \lambda(v_1, v_2, \ldots, v_{k-1}) \geq \max \mu(v_1, v_2, \ldots, v_{k-1}),$$

or what is the same thing, $\lambda_k \geq \mu_k$.

We now illustrate our theorem by an example. Let there be given two elastic plates of the same material and the same thickness which are rigidly fixed at their edges. We denote by S_1 and S_2 the areas of the plates and we suppose that S_1 fits entirely within S_2 (Fig. 12). We further denote by L_1 and L_2

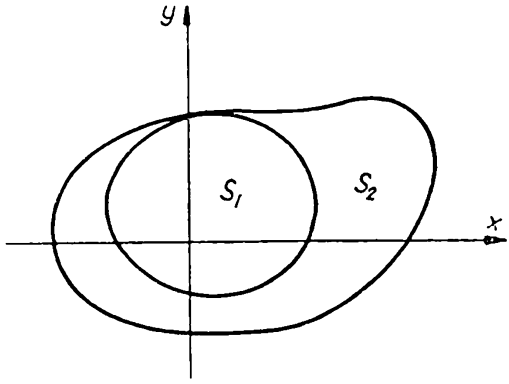

FIG. 12

the edges of the domains S_1 and S_2. Further, let A denote a biharmonic operator defined for functions which satisfy the boundary conditions

$$w\Big|_{L_1} = 0, \quad \frac{\partial w}{\partial n}\Big|_{L_1} = 0; \tag{10}$$

and similarly let B denote a biharmonic operator defined for functions which satisfy the conditions

$$w\Big|_{L_2} = 0, \quad \frac{\partial w}{\partial n}\Big|_{L_2} = 0. \tag{11}$$

Then the squares of the frequencies of the characteristic vibrations of the plates S_1 and S_2 are proportional to the eigen-

values of the operators A and B. We shall prove that $A \geq B$. We note, first of all, that

$$\|w\|_A^2 = \int_{S_1}\int \left\{\left(\frac{\partial^2 w}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 w}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 w}{\partial y^2}\right)^2\right\} dS, \tag{12}$$

$$\|w\|_B^2 = \int_{S_2}\int \left\{\left(\frac{\partial^2 w}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 w}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 w}{\partial y^2}\right)^2\right\} dS, \tag{13}$$

w satisfying conditions (10) and (11) in the first and second integrals respectively.

Let w satisfy conditions (10) and impart a finite value to integral (12). Then this function w possesses finite energy with respect to the operator A. We extrapolate the function w into S_2 by putting it equal to zero outside S_1. The function w thus defined satisfies conditions (11) and possesses finite energy with respect to operator B, where obviously $\|w\|_A^2 = \|w\|_B^2$. Corresponding to the definition given above $A \geq B$. Now from theorem 1 follows:

If two elastic plates of the same thickness and the same material are such that one of them can fit entirely within the other, and the edges of both plates are rigidly fixed, then, for any k, the k-th frequency of the characteristic vibration of the smaller plate is greater than the k-th frequency of the larger plate.

It is even simpler to prove that the frequencies of the characteristic vibrations of a plate with a rigidly fixed edge are greater than the corresponding frequencies of the characteristic vibrations of the same plate whose edge is entirely or partially supported or free.

The number of such examples can be increased very easily.

CHAPTER VI

GENERALIZATION OF THE PRECEDING RESULTS

THE READER, without doubt, will have noticed that in discussing the energy method the concrete form of the particular operator was of secondary importance, the essentials being certain general properties of the operators considered, such as their symmetry, positive definiteness or positive-bounded-below nature. In their turn these properties are closely related to the concept of scalar product which for all its simplicity is exceptionally important.

It therefore seems sensible to give the energy method a more general discussion in which all the essentials are more sharply defined while the less important details pass into the background. Such a more general formulation of the energy method is founded upon a more general concept of scalar product connected with the concept of *Hilbert space*.

The present chapter is devoted to this range of ideas. The reader who is obliged to study the general theory of operators in Hilbert space can turn to the fundamental works of V. I. Smirnov [3], N. I. Akhiyezer and I. M. Glazman [1] or F. Riesz and B. De Sz. – Nagy [1]. The application of the general theory to the differential operators of mathematical physics is gi ven in the books by S. L. Sobolev [2] and by the author [11]. We start the chapter with a brief exposition of the concept of Lebesgue integral and its fundamental properties.

The device of Lebesgue integral is scarcely useful in carrying out actual calculations. The device plays however an essential role in theoretical questions owing to the unusual "uni-

versal" nature of the Lebesgue integral. The point is that the Lebesgue integral exists for practically any bounded function, while the ordinary Riemann integral exists only when certain, sometimes intractable, conditions are imposed upon the discontinuities of the function. Further the conditions for Lebesgue integrability of an unbounded function are established far more easily than the conditions for Riemann integrability. Thus, the use of the Lebesgue integral practically rids us of the necessity of proving the integrability of the functions considered, and this considerably simplifies the investigation.

A reasonably complete exposition of the theory of Lebesgue integrals will be found by the reader who is interested in this question in the books by V. I. Smirnov [3] and I. P. Natanson [1].

§ 41. The concept of Lebesgue integral

The construction of the Lebesgue integral rests upon the concept of *the measure of a point set*. As a specific case we shall consider the set of points lying within the range $a \leq x \leq b$. If a set of points fills some interval $\alpha < x < \beta$ lying within the range $a \leq x \leq b$ then as the measure of this set we simply adopt the length $\beta - \alpha$. We now consider a set which fills a finite or enumerable ensemble of non-intersecting intervals. (An infinite set is called enumerable if all its elements can be numbered with the aid of the positive integers. Examples of enumerable sets are: the set of positive integers itself, any of its infinite parts (e.g. the set of even numbers, the set of squares of integers etc.), the set of positive and negative integers. It can be proved that the set of all rational numbers is also enumerable. Examples of sets which are not enumerable are the sets of points filling an interval, plane, or spatial domain.)

Such a set is called an open set, and the set remaining when the open set has been removed from the range $a \leq x \leq b$ is known as a closed set. It is natural to define the measure of an open set as the sum of the lengths of its component inte-

grals and the measure of a closed set as the difference between the length of the segment $a \leq x \leq b$ and the measure of the open set which has been removed.

Passing on to the general definition of measure we conventionally say that set A encloses (or includes) set B, if all the points of B belong to A (more concisely if B is a part of A). We shall describe as the outer measure of set A, the exact lower bound of the measures of the open sets enclosed by it, and as its inner measure, the exact upper bound of the closed sets enclosed by set A. It can be proved (intuitively this is obvious) that the inner measure of a set does not exceed its outer measure. Further a set is called measurable if its outer and inner measures coincide.

The common inner and outer measures of a measurable set is called its *Lebesgue measure*. We shall denote the measure of set A by the symbol μA. Measurability is a considerably more widely distributed property of sets than it may appear at first sight.

Effectively all the sets constructed until now are measurable. At present we have succeeded in proving the existence of immeasurable sets only by leaning on the so called selection principle which in effect does not enable us to construct a set.

In what follows, when speaking of a point set, we shall always assume it to be measurable.

The measure of a set possesses a number of properties which are inherent in length (the concept of measure being a generalization of the concept of length). We define *the sum* of two sets as that third set into which all the elements, both of the first and the second set, enter, common elements belonging to both component sets entering only once into the sum. When two sets do not intersect (i.e. do not have common points) then the measure of the sum of the sets equals the sum of the measures of the components; the measure of the sum of two intersecting sets equals the sum of their measures minus the measure of their common part, and so on.

The concept of measure can be extended in natural fashion to point sets distributed in a plane or in three dimensional space. Here the concept of measure is naturally the generalized concept of area or volume. One can also readily extend the concept of measure to point sets in multidimensional spaces.

There exist sets with measure equal to zero. These obviously are sets which can be enclosed by open sets of as small a measure as required. Finite and countable sets of points are of this type, as can readily be proved. There also exist uncountable sets of measure zero. It is commonly said that a property occurs *almost everywhere* if it cannot occur only for a set of measure zero.

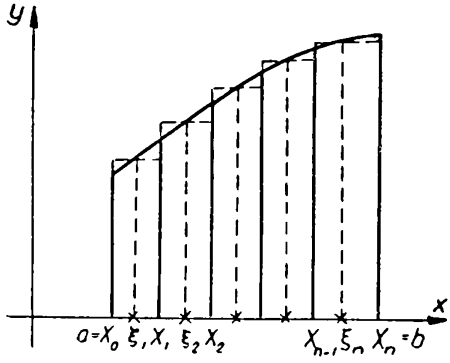

FIG. 13

We now turn to the concept of integral. We recall that by the definition due to Cauchy and Riemann, a definite integral is considered as the limit of the integral sum (the meaning of the symbols being given in Fig.13)

$$\int_a^b f(x)\mathrm{d}x = \lim_{\lambda \to 0} \sum_{k=1}^{n} f(\xi_k)(x_k - x_{k-1}), \tag{1}$$

where λ is the largest of the lengths of the intervals (x_{k-1}, x_k). If we take a sufficiently fine subdivision of the basic range

$a \leq x \leq b$ and drop the limit sign on the right in (1), we obtain an approximate equation

$$\int_a^b f(x)\mathrm{d}x \approx \sum_{k=1}^{n} f(\xi_k)(x_k - x_{k-1}), \tag{2}$$

which corresponds to the following geometric model: the area of the thin strip bounded by the arc of the curve $y=f(x)$, the segment (x_{k-1}, x_k) of the abscissa axis and the two ordinates drawn through the points x_{k-1} and x_k, isreplaced by the area of a rectangle whose base is the segment (x_{k-1}, x_k) and whose height is the ordinate of the curve drawun throgh an arbitrarily selected point ξ_k in the given segment. The error in the approximate equation (2) is small if the ordinate of the curve $y=f(x)$ does not change rapidly; in this case the particular choice of the point ξ_k is virtually without effect upon the magnitude of the error.

However, formula (2) becomes unreliable if the function $f(x)$ oscillates rapidly (Fig. 14) even though it remains continuous. Obviously we will obtain entirely different results depending on whether we take as the point ξ_k the greatest, least or some intermediate ordinate. In this case it is expedient to evaluate the area in another way. We lay off on the ordinate axis the maximum and minimum values of the function (or more precisely the upper and lower bounds of the values of the function). Let these be the points α and β.

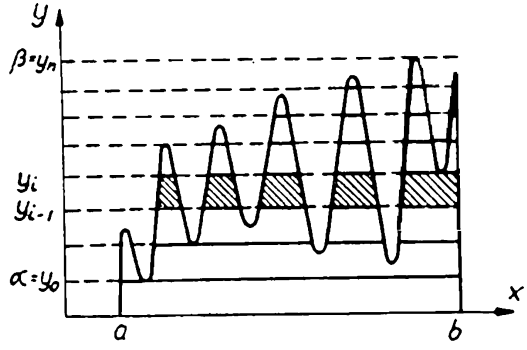

FIG. 14

We divide the range $\alpha \leq y \leq \beta$ into small portions by inserting the points of subdivision $y_0 = \alpha, y_1, y_2, \ldots, y_n = \beta$. Through the points of subdivision we draw straight lines parallel to the abscissa axis. The area of interest to us is represented by shapes reminiscent of rectangles of small height; these "rectangles" are distributed in horizontal strips between the lines $y = y_{i-1}$ and $y = y_i$, $i = 1, 2, \ldots, n$. We estimate the sum of the areas of the "rectangles" included in such a strip by taking for example the side of the rectangle lying on the upper side of the strip ($y = y_i$) as the base of the "rectangle". The sum of the areas obtained thus approximately equals the product of the height $y_i - y_{i-1}$ and the sum of the lengths of the bases. Obviously at each point on the base the ordinate $f(x) \geq y_i$. Thus the sum of the bases of the "rectangles" can be characterized by the set of values x at which $f(x) \geq y_i$, and the sum of the lengths of the bases, by the measure of this set. (It stands to reason that we assume the measurability of this set for any y_i.) We denote the measure obtained thus by μ_i. Then the area of that part of the figure which is included in the strip $y_{i-1} < y < y_i$ (this part is hatched in Fig. 14) approximately equals $\mu_i(y_i - y_{i-1})$, and the area of the entire figure approximately equals $\mu_0 y_0 + \sum_{i=1}^{n} \mu_i(y_i - y_{i-1})$.

It is interesting to note that the precision of this approximate equation depends little upon whether the function $f(x)$ is continuous or discontinuous providing it remains bounded. We shall not stop to prove this assertion; we merely note that its meaning is completely explained by the comparison of Figs. 14 and 15 in the first of which a graph of a continuous function is depicted and in the second, a graph of a function having discontinuities at the points c_1, c_2, c_3. The approximate evaluation of the area as the sum

$$\mu_0 y_0 + \sum_{i-1}^{n} \mu_i(y_i - y_{i-1}) \tag{3}$$

in both cases gives an error of the same order.

We shall now decrease the intervals (y_{i-1}, y_i), so that the length of the greatest of them tends to zero. If the sum (3) then tends to some limit then this limit is known as the *Lebesgue integral* of the function $f(x)$ over the segment $a \leq x \leq b$ Generalizing this construction for the case of functions which are not necessarily continuous, we arrive at the following definition of Lebesgue integral:

The limit

$$\lim_{\lambda \to 0} \left\{ \mu_0 y_0 + \sum_{i=1}^{n} \mu_i (y_i - y_{i-1}) \right\}, \qquad \lambda = \max(y_i - y_{i-1}),$$

where μ_i is the measure of the set of values x for which $f(x) \geq y$ and $\alpha = y_0$ and $\beta = y_n$ are the numbers between which all the values of $f(x)$ are included is called the Lebesgue integral of the bounded function $f(x)$ over the range $a \leq x \leq b$.

The Lebesgue integral of a function in two or more independent variables is defined in a completely analogous fashion.

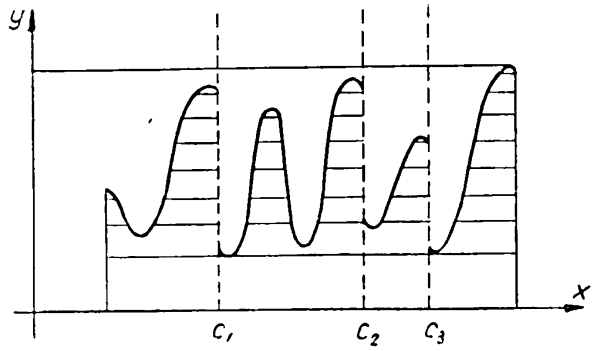

FIG. 15

It can be proved that the Lebesgue integral of a bounded function always exists. We assume, of course, that for any value of the constant A the set of values of x for which $f(x) \geq A$ is measurable – otherwise the definition of the Lebesgue integral loses meaning. In fact for all the functions

which we have been able to construct until now the sets indicated are measurable.

If there exists a definite integral in the ordinary (Riemannian) sense, then, as can be proved, it coincides with the Lebesgue integral. Therefore there is no necessity for a special symbol to denote a Lebesgue integral, and we shall use the ordinary symbols

$$\int_a^b f(x)\mathrm{d}x, \quad \int_\Omega\int f(x, y)\mathrm{d}x\mathrm{d}y,$$

etc.

We now give a definition of Lebesgue integral for a non-bounded function. We first of all suppose that the function $f(x)$ is non-negative. We denote by N an arbitrary positive number and we introduce the new function $f_N(x)$ by putting (see Fig. 16)

$$f_N(x) = \begin{cases} f(x), & f(x) \leq N \\ N, & f(x) > N \end{cases}$$

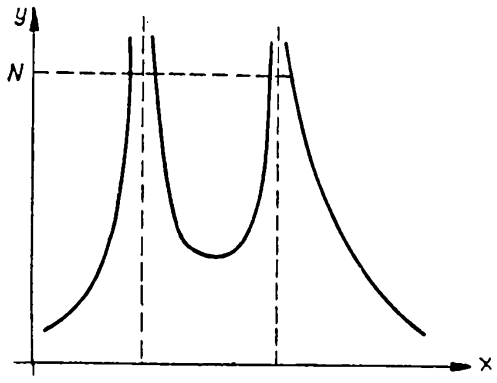

FIG. 16

The function $f_N(x)$ is bounded since, obviously, $0 \leq f_N(x) + N$ therefore its Lebesgue integral

$$\int_a^b f_N(x)\mathrm{d}x \tag{4}$$

exists. This integral increases (or in the extreme case does not decrease) with increasing N and therefore tends with indefinitely increasing N to a definite, finite or infinite limit. *If there exists a finite limit of integral* (4) *as* $N \to \infty$, *then this limit is called the Lebesgue integral of the non-negative, unbounded function* $f(x)$. Thus, by definition, if the function $f(x) \geq 0$ and unbounded, then

$$\int_a^b f(x)\mathrm{d}x = \lim_{N\to\infty} \int_a^b f_N(x)\mathrm{d}x.$$

An unbounded, non-negative function, for which the Lebesgue integral in the range from a to b exists is called *summable* over the range $a \leq x \leq b$. Functions which are summable in an m-dimensional domain ($m \geq 2$) are analogously defined.

Now let $f(x)$ have either sign. It can be represented as the difference of two non-negative functions. In fact, if we put

$$f_1(x) = \frac{1}{2}\{|f(x)| + f(x)\}, \quad f_2(x) = \frac{1}{2}\{|f(x)| - f(x)\},$$

then $f_1(x) \geq 0$, $f_2(x) \geq 0$ and $f(x) = f_1(x) - f_2(x)$. *We shall consider* $f(x)$ *summable if and only if each of the functions* $f_1(x)$ *and* $f_2(x)$ *are summable and in this case we put by definition*

$$\int_a^b f(x)\mathrm{d}x = \int_a^b f_1(x)\mathrm{d}x - \int_a^b f_2(x)\mathrm{d}x.$$

If $f(x)$ is summable the function $|f(x)| = f_1(x) + f_2(x)$ is also summable. Thus the Lebesgue integral is always absolutely convergent.

If the function $f(x)$ is complex, $f(x) = \varphi(x) + i\psi(x)$ then we assume

$$\int_a^b f(x)\mathrm{d}x = \int_a^b \varphi(x)\mathrm{d}x + i\int_a^b \psi(x)\mathrm{d}x.$$

By definition the integral on the left exists if and only if both integrals on the right exist. The Lebesgue integral of a complex function is also absolutely convergent.

The fundamental properties of a Riemann integral still hold true for the Lebesgue integral. We also note three properties which are characteristic for a Lebesgue integral and play an important role in all the following. The reader will find their proofs in the courses cited at the beginning of the chapter.

1. The value of a Lebesgue integral is unchanged if the value of the integrand function is changed over a set of measure zero. Moreover, a Lebesgue integral retains its meaning if the value of the function remains undefined for a set of measure zero. Consequently, it follows that a Lebesgue integral equals zero if the integrand function is almost everywhere equal to zero.

2. If the integral of a non-negative function equals zero, then this function equals zero almost everywhere.

3. If $|f(x) \leq \varphi(x)$ and $\varphi(x)$ is summable, then $f(x)$ is also summable.

The proof is as follows. We construct the function $|f(x)|_N$ as indicated above. Being bounded it is integrable, so

$$\int_a^b |f(x)|_N dx \leq \int_a^b \varphi(x) dx, \qquad (*)$$

since, obviously $|f(x)|_N \leq |f(x) \leq \varphi(x)$. The integral on the left in (*) increases monotonically with increasing N. Being bounded by virtue of inequality (*), this integral has a finite limit. But then, by definition, $f(x)$ is summable.

We note that the Riesz–Fischer theorem formulated in § 16 about convergence in the mean asserts about the limiting function only that its square is Lebesgue integrable, even in the case where the squares of all the functions of the sequence are integrable in the usual sense of the word (i.e. in the Riemann sense cf. 1, **116–117**. We) shall further show that the Riesz–Fischer theorem also holds true when the functions of the sequence assume complex values. We now present a formulation of this theorem applicable to the given case.

We consider a family of functions which in general assume complex values which are defined almost everywhere in some finite domain Ω of m-dimensional Euclidian space and which possess the property that their squares are summable in Ω. For brevity we shall also call the functions *quadratically summable.*

We here, in what follows, term two functions coincident if they coincide almost everywhere in Ω.

We shall say that a sequence of functions $\varphi_n(P)$ which are quadratically summable in Ω *converges in the mean* to a function $\varphi(P)$ which is quadratically summable in Ω, if

$$\lim_{n\to\infty}\int_{\Omega}|\varphi_n(P)-\varphi(P)|^2\mathrm{d}\Omega=0.$$

The Riesz–Fischer Theorem. If the functions $\varphi_n(P)$, $n=1, 2, \ldots$ *which are quadratically summable in* Ω, *satisfy the condition*

$$\lim_{k,\,n\to\infty}\int_{\Omega}|\varphi_k(P)-\varphi_n(P)|^2\mathrm{d}\Omega=0,$$

then there exists a function $\varphi(P)$ *which is quadratically summable in* Ω *to which thesequence* $\varphi_n(P)$ *converges in the mean.*

§ 42. Functional Hilbert spaces

We shall consider a set of functions defined in the finite domain Ω. Bearing in mind certain later applications, we shall consider that these functions, in general, assume complex values. The concept of linear set presented in § 4 may be extended for such functions: a set of functions is called linear if together with the functions $\varphi(P)$ and $\psi(P)$ it also contains the functions $\varphi(P)+\psi(P)$ and $a\varphi(P)$ where a is any (this time in general, complex) constant. For the following it is important to note that a set of functions quadratically summable in Ω is also linear; if $\varphi(P)$ and $\psi(P)$ are quadratically summable then from property 3 of Lebesgue integrals (§ 41)

and from the inequality $|a\varphi+b\psi|^2 \leq \frac{1}{2}\{|a|^2|\varphi|^2+|b|^2|\psi|^2\}$ it follows that the function $a\varphi(P)+b\psi(P)$ is also quadratically summable.

A linear set is called a functional Hilbert space if for each pair of functions $\varphi(P)$, $\psi(P)$ belonging to the linear set a number (φ, ψ) is generated which satisfies the following axioms:

A. $(\varphi, \psi)=\overline{(\psi, \varphi)}$.

B. $(a_1\varphi_1+a_2\varphi_2, \psi)=a_1(\varphi_1, \psi)+a_2(\varphi_2, \psi)$,

where φ_1, φ_2, ψ are elements of the linear set and a_1 and a_2 are constants. (The upper bar denotes the complex conjugate.)

C. $(\varphi, \varphi)\geq 0$.

D. If $(\varphi, \varphi)=0$, then $\varphi(P)\equiv 0$.

We shall call the functions which enter into the Hilbert space the elements (and sometimes, the points) of this space. We shall call the number (φ, ψ) *the scalar product* of the elements φ and ψ.

We direct the reader's attention to the fact that property A of scalar multiplication does not coincide with the analogous property established in § 3. We will explain the reason for this a little later.

Besides the "complex" space that we have defined it is possible to consider "real" functional Hilbert spaces. Their elements are real functions the scalar multiplication of which possesses properties A–D of § 3, so that in particular, a scalar product is symmetric $(\varphi, \psi)=(\psi, \varphi)$. In the preceding chapters we have, in essence, been concerned with real Hilbert spaces.

We now establish simple properties of scalar product, similar (but not always identical), to those presented in § 3.

From axioms A and B it follows that $(\varphi, \lambda\psi)=\bar{\lambda}(\varphi, \psi)$ where λ is a complex number. In fact by axiom A, $(\varphi, \lambda\psi)=\overline{(\lambda\psi, \varphi)}$. By virtue of axiom B, $\overline{(\lambda\psi, \varphi)}=\bar{\lambda}\ \overline{(\psi, \varphi)}$ which by virtue of axiom A equals $\bar{\lambda}(\varphi, \psi)$. Finally $(\varphi, \lambda\psi)=\bar{\lambda}(\varphi, \psi)$.

It is now clear that a scalar product whose factors are sums can be composed according to the rule for the multiplication of polynomials with the one difference, however, that the numerical coefficient of the second factor must be replaced by the complex conjugate quantity in carrying it outside the sign of scalar product. It becomes clear from what has been said why it proved necessary to change property A of scalar product. In fact, by property C, we must have not only $(\varphi, \varphi) \geq 0$ but also $(\lambda\varphi, \lambda\varphi) \geq 0$, where λ is any complex number. If we were to retain property A in the form $(\varphi, \psi) = (\psi, \varphi)$ then we would arrive at the fact that $(\lambda\varphi, \lambda\varphi) = \lambda^2 (\varphi, \varphi)$ which for arbitrary complex λ would be a complex quantity and not positive or equal to zero.

The non-negative number $\sqrt{(\varphi, \varphi)}$ is called the *norm* of the element φ and is denoted by the symbol $\|\varphi\|$, so that

$$\|\varphi\| = \sqrt{(\varphi, \varphi)};$$

the norm of the difference of two elements is called the *distance* between them. We shall say that formula (1) defines the *metric* in Hilbert space.

An element of Hilbert space whose norm equals unity is called *normalized*.

The norm in Hilbert space possesses properties (a)–(d), established in § 3. Properties (a) and (b) are obvious, properties (c) Cauchy's inequality, and (d) the triangle inequality require a fresh proof.

Thus, we shall prove that for any elements φ and ψ of Hilbert space Cauchy's inequality

$$|(\varphi, \psi)| \leq \|\varphi\| \, \|\psi\| \tag{2}$$

is valid. This inequality is obvious if $(\varphi, \psi) = 0$. thus it suffices to consider the case where $(\varphi, \psi) \neq 0$.

Let θ be a complex number which we shall fix below and let λ be an arbitrary real number. By axiom C,

$$(\varphi - \lambda\bar{\theta}\psi, \varphi - \lambda\bar{\theta}\psi) \geq 0.$$

We expand brackets, at the same time placing the numerical coefficients outside the scalar product:

$$(\varphi, \varphi)-\lambda\theta(\varphi, \psi)-\lambda\bar{\theta}(\psi, \varphi)+\lambda^2|\theta|^2(\psi, \psi)\geq 0,$$

or, what is the same thing,

$$\|\varphi\|^2-2\lambda\mathrm{Re}[\theta(\varphi, \psi)]+\lambda^2|\theta|^2\|\psi\|^2\geq 0.$$

(Here, as everywhere in what follows, we denote by the symbols Re and Im the real and imaginary parts of the expression immediately following.)

The expression on the left is quadratic in λ and is non-negative for all real values of the argument. But in such a case its discriminant is non-positive

$$\{\mathrm{Re}[\theta(\varphi, \psi)]\}^2-|\theta|^2\|\varphi\|^2\|\psi\|^2\leq 0.$$

Hence, taking the root, we find

$$|\mathrm{Re}[\theta(\varphi, \psi)]|\leq|\theta|\,\|\varphi\|\,\|\psi\|. \tag{3}$$

In this inequality, which is valid for any θ, we put

$$\theta=\frac{|(\varphi, \psi)|}{(\varphi, \psi)}.$$

This is possible, since $(\varphi, \psi)\neq 0$ by assumption.

With such a choice of θ we have $|\theta|=1$. Further $\theta(\varphi, \psi)=|(\varphi, \psi)|$ is a real (in fact positive) quantity and thus coincides with its real part: $\mathrm{Re}[\theta(\varphi, \psi)]=\theta(\varphi, \psi)=|(\varphi, \psi)|$. Substituting this in (3) we arrive at inequality (2).

The triangle inequality is also readily established

$$\|\varphi+\psi\|\leq\|\varphi\|+\|\psi\|. \tag{4}$$

We have

$$\begin{aligned}\|\varphi+\psi\|^2\leq(\varphi+\psi, \varphi+\psi)&=(\varphi, \varphi)+(\varphi, \psi)+(\psi, \varphi)+(\psi, \psi)\\&=\|\varphi\|^2+2\mathrm{Re}[(\varphi, \psi)]+\|\psi\|^2.\end{aligned}$$

By Cauchy's formula $|(\varphi, \psi)|\leq\|\varphi\|\,\|\psi\|$. *A fortiori* $|\mathrm{Re}[(\varphi, \psi)]|\leq\|\varphi\|\,\|\psi\|$. Hence

$$\|\varphi+\psi\|^2\leq\|\varphi\|^2+2\|\varphi\|\,\|\psi\|+\|\psi\|^2,$$

which is equivalent to the inequality (4).

We now present a few examples of functional Hilbert spaces.

(a) *The space* $L_2(\Omega)$. As was noted at the beginning of the section a set of functions defined almost everywhere in Ω and quadratically summable in Ω is linear. We define the scalar product for this linear set, by putting

$$(\varphi, \psi) = \int_{\Omega} \varphi(P)\overline{\psi(P)}\,d\Omega. \tag{5}$$

Integral (5) exists, as follows from property 3 § 41 and from the inequality $|\varphi\bar{\psi}| \leq \frac{1}{2}\{|\varphi|^2 + |\psi|^2\}$. Expression (5) obviously satisfies axioms A–C; it will also satisfy axiom D, if as we agreed in § 41, we consider two functions to be equal which are equal almost everywhere, and, in particular, if we do not distinguish from the identically zero function a function which equals zero almost everywhere.

By introducing the scalar product we have reduced our linear set to a Hilbert space; it is usually denoted by the symbol $L_2(\Omega)$. If Ω is the part $a \leq x \leq b$ of a straight line then one writes $L_2(a, b)$. The norm in $L_2(\Omega)$ is defined by the equation

$$\|\varphi\|^2 = \int_{\Omega} |\varphi^2(P)|\,d\Omega. \tag{6}$$

Inequality (2) here transforms into the well-known Cauchy inequality,

$$\left|\int_{\Omega} \varphi(P)\overline{\psi(P)}\,d\Omega\right|^2 \leq \int_{\Omega} |\varphi^2(P)\,d\Omega \int_{\Omega} |\psi^2(P)|\,d\Omega.$$

The importance of space $L_2(\Omega)$ in all the foregoing is obvious; it will be of no less importance in the succeeding sections. Perhaps it merely remains to note that earlier we have used the real space $L_2(\Omega)$, the elements of which were real quadratically summable functions; above we have called them "functions with finite norm".

(b) Another important example of a functional Hilbert space

is the space $L_2(\Omega;\sigma(P))$. Its elements are functions for which the integral

$$\int_{\Omega}\sigma(P)|\varphi^2(P)|\,d\Omega \tag{7}$$

exists. Here $\sigma(P)$ is some non-negative function which is the same for all the elements of the space. In this case we consider two functions $\varphi_1(P)$ and $\varphi_2(P)$ to be identical if

$$\int_{\Omega}\sigma(P)|\varphi_1(P)-\varphi_2(P)|^2 d\Omega=0.$$

The scalar product in $L_2(\Omega;\sigma(P))$ is specified by the formula

$$(\varphi,\ \psi)=\int_{\Omega}\sigma(P)\varphi(P)\overline{\psi(P)}\,d\Omega, \tag{8}$$

and the norm, as follows from (8), by the formula

$$\|\varphi\|^2=\int_{\Omega}\sigma(P)|\varphi^2(P)|\,d\Omega. \tag{9}$$

(c) We denote by S the boundary of the domain Ω and consider the set of functions which are continuous in the closed domain $\bar{\Omega}=\Omega+S$ together with all their derivatives up to order r inclusive.

It is easily seen that this set is linear. We shall reduce it to a Hilbert space if we specify the scalar product in it by the following formula:

$$(\varphi,\ \psi)=\sum_{k=0}^{r}\sum_{\alpha_1+\alpha_2+\ldots+\alpha_m=k}\int_{\Omega}\frac{\partial^k\varphi}{\partial x_1^{\alpha_1}\partial x_2^{\alpha_2}\ldots\partial x_m^{\alpha_m}}\times\frac{\partial^k\bar{\psi}}{\partial x_1^{\alpha_1}\partial x_2^{\alpha_2}\ldots\partial x_m^{\alpha_m}}\,d\Omega. \tag{10}$$

Here $x_1, x_2, \ldots, x_m$ are the cartesian coordinates of a variable point in Ω. This space is denoted by the symbol $L_2^{(r)}(\Omega)$; for $r=0$ it transforms to the space $L_2(\Omega)$. As is seen from formula (10), the norm in $L_2^{(r)}(\Omega)$ is defined by the equation

$$\|\varphi\|^2=\sum_{k=0}^{r}\sum_{\alpha_1+\alpha_2+\ldots+\alpha_m=k}\int_{\Omega}\left|\frac{\partial^k\varphi}{\partial x_1^{\alpha_1}\partial x_2^{\alpha_2}\ldots\partial x_m^{\alpha_m}}\right|^2 d\Omega. \tag{11}$$

We now refer to another class of Hilbert spaces which are of importance in connexion with previous chapters. Let A be a positive definite operator. The set of functions (we consider them to be real) which constitute the field of definition D_A of operator A can be made a real Hilbert space if we take as the scalar product of two such functions their energy product:

$$[u, v]=(Au, v);\ u\in D_A,\ v\in D_A.$$

The norm in such a space is equivalent to the energy norm defined in Ch. II. We shall most frequently denote the Hilbert space thus constructed by H_0; if the necessity of stressing the role of operator A arises then we shall use the symbol H_A to designate this space.

Sometimes there occur Hilbert spaces of a more general nature, for instance, those whose elements are not functions but a sequence of numbers. We restrict ourselves here to a mention of the Hilbert space of vector functions (in general complex) which are important in many applications and which we shall denote by $\mathbf{L}_2(\Omega)$.

Its elements are vector point functions of the domain Ω; if $\boldsymbol{\varphi}(P)$ is such a vector with components along the coordinate axes φ_x, φ_y, φ_z, then we require that the integral

$$\int_\Omega \{|\varphi_x|^2+|\varphi_y|^2+|\varphi_z|^2\}\mathrm{d}\Omega$$

exists. Scalar product in $\mathbf{L}_2(\Omega)$ is defined by the formula

$$(\boldsymbol{\varphi}, \boldsymbol{\psi})=\int_\Omega (\varphi_x\overline{\psi}_x+\varphi_y\overline{\psi}_y+\varphi_z\overline{\psi}_z)\,\mathrm{d}\Omega. \tag{12}$$

The norm of element $\boldsymbol{\varphi}$, obviously, equals

$$\|\boldsymbol{\varphi}\|^2=\int_\Omega \left\{|\varphi_x|^2+|\varphi_y|^2+|\varphi_z|^2\right\}\mathrm{d}\Omega. \tag{13}$$

If the vector functions considered are real, then we obtain the real vector Hilbert space $\mathbf{L}_2(\Omega)$; the scalar product and norm therein are defined by the simpler formulae

$$(\boldsymbol{\varphi}, \boldsymbol{\psi})=\int_\Omega \boldsymbol{\varphi}\cdot\boldsymbol{\psi}\mathrm{d}\Omega;\ \ \|\boldsymbol{\varphi}\|^2=\int_\Omega \boldsymbol{\varphi}^2\mathrm{d}\Omega,$$

where $\boldsymbol{\varphi}.\boldsymbol{\psi}$ denotes the ordinary scalar product of vectors, and φ^2 denotes the square of the length of the vector $\boldsymbol{\varphi}$. We essentially made use of the vector space $\mathbf{L}_2(\Omega)$ in the problems of elasticity theory; as we shall see in the next chapter, this space also proves useful in other problems.

We now extend the definition of linear dependence given in § 9 to a Hilbert space. The elements $\varphi_1, \varphi_2, \ldots, \varphi_n$ of the Hilbert space are known as *linearly dependent* if it is possible to find constant numbers $a_1, a_2, \ldots, a_n$ not all equal to zero, such that the following equation is satisfied.

$$a_1\varphi_1 + a_2\varphi_2 + \ldots + a_n\varphi_n = 0.$$

These elements are termed *linearly independent* if the latter equation is satisfied only for

$$a_1 = a_2 = \ldots = a_n = 0.$$

A theorem completely analogous to theorem 2 § 9 is valid.

THEOREM. *In order that the elements $\varphi_1, \varphi_2, \ldots, \varphi_n$ of Hilbert space be linearly dependent it is necessary and sufficient that the determinant*

$$\begin{vmatrix} (\varphi_1, \varphi_1), (\varphi_1, \varphi_2), \ldots, (\varphi_1, \varphi_n) \\ (\varphi_2, \varphi_1), (\varphi_2, \varphi_2), \ldots, (\varphi_2, \varphi_n) \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ (\varphi_n, \varphi_1), (\varphi_n, \varphi_2), \ldots, (\varphi_n, \varphi_n) \end{vmatrix} \tag{14}$$

equal zero.

The proof given in § 9 extends to the general case without change.

As before, we shall call determinant (14) the Gram determinant of the elements $\varphi_1, \varphi_2, \varphi_3, \ldots, \varphi_n$.

§ 43. The limit process in Hilbert spaces

Let there be given a Hilbert space H and let $\varphi_1, \varphi_2, \ldots, \varphi_n$ be a sequence (which we shall often denote by the symbol $\{\varphi_n\}$) of its elements. We shall say that this sequence converges, or tends, to φ if φ is an element of H and

$$\lim_{n\to\infty} \|\varphi_n - \varphi\| = 0. \tag{1}$$

The element φ is called *the limit* of sequence $\{\varphi_n\}$.
We shall denote this circumstance by using the usual symbol

$$\varphi_n \to \varphi, \quad \lim_{n\to\infty} \varphi_n = \varphi.$$

We shall clarify our definition by examples.
(a) In the space $L_2(\Omega)$ the convergence $\varphi_n \to \varphi$ means that

$$\lim_{n\to\infty} \int_{\Omega} |\varphi_n(P) - \varphi(P)|^2 d\Omega = 0,$$

i.e. that $\varphi_n(P)$ converges to $\varphi(P)$ in the mean.
(b) In the same way convergence in $L_2(\Omega; \sigma(P))$ is convergence in the mean with the weight function $\sigma(P)$.
(c) Convergence in $L_2^{(r)}(\Omega)$, as is easily seen, denotes convergence in the mean of the functions of the sequence and of all their derivatives up to r inclusive to a limiting function and its corresponding derivatives.
(d) Convergence in the space H_A (see § 42) is convergence in the energy of operator A.

The theorem which is a generalization of theorem 2 § 7 is valid:

THEOREM 1. *If* $\varphi_n \to \varphi$ *and* $\psi_n \to \psi$ *then*

$$(\varphi_n, \psi_n) \to (\varphi, \psi).$$

As is readily verified,

$$(\varphi_n, \psi_n) - (\varphi, \psi) = (\varphi_n - \varphi, \psi_n - \psi) + (\varphi, \psi_n - \psi) + (\varphi_n - \varphi, \psi).$$

Hence Cauchy's formula,

$$|(\varphi_n, \psi_n) - (\varphi, \psi)| \leq \|\varphi_n - \varphi\| \, \|\psi_n - \psi\| + \|\varphi\| \, \|\psi_n - \psi\| + \|\varphi_n - \varphi\| \, \|\psi\| \to 0,$$

which proves the theorem.

COROLLARY 1. If $\varphi_n \to \varphi$ then $(\varphi_n, \psi) \to (\varphi, \psi)$.

COROLLARY 2. If $\varphi_n \to \varphi$ then $\|\varphi_n\| \to \|\varphi\|$.

We obtain corollary 1 from the theorem by putting $\psi_n = \psi$, and corollary 2, by putting $\psi_n = \varphi_n$.

We note that corollary 2 is the direct generalization of theorem 2 § 7 for the case of an arbitrary Hilbert space.

We now turn to the general concept of convergence. Let there be specified in a Hilbert space, which we shall denote by H, a sequence $\{\varphi_n\}$ which converges to the element φ. By definition, $\lim_{n\to\infty} \|\varphi_n - \varphi\| = 0$. This means that for a specified $\varepsilon > 0$ it is possible to find $n_0(\varepsilon)$ such that for $n \geq n_0(\varepsilon)$, $\|\varphi_n - \varphi\| < \varepsilon$. Let $k \geq n_0\left(\frac{\varepsilon}{2}\right)$ and $n \geq n_0\left(\frac{\varepsilon}{2}\right)$. Then

$$\|\varphi_k - \varphi\| < \frac{\varepsilon}{2} \text{ and } \|\varphi_n - \varphi\| < \frac{\varepsilon}{2}.$$

We shall estimate the norm of the difference $\|\varphi_k - \varphi_n\|$. By the triangle inequality,

$$\|\varphi_k - \varphi_n\| = \|(\varphi_k - \varphi) - (\varphi_n - \varphi)\| \leq \|\varphi_k - \varphi\| + \|\varphi_n - \varphi\| < \varepsilon.$$

The last equation, by virtue of the arbitrariness of ε, is equivalent to the equation

$$\lim_{\substack{k\to\infty \\ n\to\infty}} \|\varphi_k - \varphi_n\| = 0. \tag{2}$$

Thus, if $\varphi_n \to \varphi$, then equation (2) is necessarily satisfied. The question arises of whether the converse assertion is true, i.e. of whether we can assert the existence of a limit of the sequence $\{\varphi_n\}$ if it satisfies equation (2). In general the reply to this question must be in the negative.

In order to clarify this we consider an example. We consider a linear set composed of functions $\varphi(x)$ which are continuous in the range $-1 \leq x \leq 1$. We reduce this linear set to

a Hilbert space by specifying the scalar product according to the formula

$$(\varphi, \psi) = \int_{-1}^{1} \varphi(x)\overline{\psi(x)}\,dx.$$

The space that we have constructed, which we denote by N is not equivalent to $L_2(-1, 1)$, since the elements of the latter can be not only continuous but also discontinuous functions.

We take a sequence of functions

$$\varphi_n(x) = \begin{cases} -1, & -1 \leq x \leq -\dfrac{1}{n}, \\ nx, & -\dfrac{1}{n} \leq x \leq \dfrac{1}{n}, \\ 1, & \dfrac{1}{n} \leq x \leq 1. \end{cases}$$

Such a function is plotted in Fig. 17. Being continuous the functions belong to N. Simple calculations show that

$$\lim_{\substack{k\to\infty \\ n\to\infty}} \|\varphi_k - \varphi_n\|^2 = \lim_{\substack{k\to\infty \\ n\to\infty}} \int_{-1}^{1} [\varphi_k(x) - \varphi_n(x)]^2 dx = 0.$$

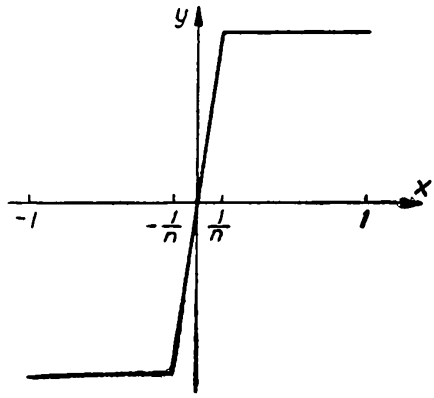

FIG. 17

The sequence $\{\varphi_n(x)\}$ tends, in the usual sense of this word, to a discontinuous function which we shall denote by $\varphi_0(x)$:

$$\varphi_0(x)=\begin{cases} -1, & -1\leq x<0, \\ 0, & x=0, \\ 1, & 0<x\leq 1, \end{cases}$$

the tending to the limit being similar in each of the segments $-1\leq x\leq -\varepsilon$ and $\varepsilon\leq x\leq 1$, where ε is any positive number. Hence it is easily seen that

$$\lim_{n\to\infty}\int_{-1}^{+1}[\varphi_n(x)-\varphi_0(x)]^2\mathrm{d}x=0,$$

i.e. that in the mean $\varphi_n(x)\to\varphi_0(x)$. The function $\varphi_0(x)$ is discontinuous, moreover, it is impossible to make it continuous by changing its value merely by a set of measure zero. Because a limit in the sense of convergence in the mean is unique, no limiting element for $\{\varphi_n(x)\}$ exists in N even though this sequence satisfies condition (2). On the other hand in the space $L_2(\Omega)$ equation (2) necessitates the existence of a limiting element – this immediately follows from the Riesz–Fischer theorem (§ 41). In our example this limiting element is the function $\varphi_0(x)$.

A Hilbert space is called complete if any sequence of its elements satisfying condition (2), *has a limit*, otherwise it is called *incomplete*. From what has been said above it follows that the space N is incomplete and $L_2(\Omega)$ is complete. It can also be proved that the space $L_2(\Omega;\sigma(P))$ is complete.

An incomplete Hilbert space can be made complete by adding certain new elements to it just as the set of rational numbers can be supplemented to that of all the (real) numbers by the introduction of the irrational numbers. Let a sequence $\{\varphi_n\}$ satisfy condition (2) but not have a limit. We attach to the sequence $\{\varphi_n\}$ as a limit the new element φ, known as the *limiting* (or *ideal*) element.

Let the limiting elements φ and ψ be defined by the sequences $\{\varphi_n\}$ and $\{\psi_n\}$ respectively. We shall assume that $\varphi = \psi$ if

$$\lim_{n\to\infty} \|\varphi_n - \psi_n\| = 0.$$

The scalar product of the limiting elements is defined by the formula

$$(\varphi, \psi) = \lim_{n\to\infty} (\varphi_n, \psi_n); \tag{3}$$

then, obviously,

$$\|\varphi\| = \lim_{n\to\infty} \|\varphi_n\|; \tag{4}$$

the existence of both limits is readily proved. It can be proved that the addition to H of all the limiting elements makes H a complete space. The actual nature of the limiting elements depends upon the completed space. Thus, the completion of space N leads to $L_2(-1, 1)$; the limiting elements here are discontinuous functions which are quadratically summable over the range $-1 \leq x \leq 1$. The completion of the space H_A leads to functions with finite energy with respect to operator A.

The completion of the space $L_2^{(r)}(\Omega)$ is connected with the concept of generalized partial derivatives; we cannot dwell on this here in greater detail.

In what follows we shall assume that a Hilbert space is complete if nothing is said to the contrary.

We introduce the following concept which will be of importance in the sequel. We shall call the set M *dense* in the Hilbert space H if each element $\varphi \in H$ can be obtained as the limit of a sequence $\varphi_n \in M$. Thus, it can be proved that the set of functions which are continuous in $\overline{\Omega} = \Omega + S$ is dense in $L_2(\Omega)$.

If a space H is obtained by the completion of an incomplete space H', then H' is dense in H. In fact, let $\varphi \in H$. If moreover, $\varphi \in H'$, then it is possible to put $\varphi_n \equiv \varphi$ and obviously $\varphi = \lim \varphi_n$ where $\varphi_n \in H'$. If $\varphi \notin H'$, then φ is a limiting element

and, by the very definition, there exists a sequence $\varphi_n \in H'$ such that

$$\varphi = \lim \varphi_n.$$

In the sequel we shall call any sequence which satisfies condition (2) *fundamental*. In a complete space, consequently, any fundamental sequence converges.

§ 44. Generalization of the concept of orthogonality

In § 10 we introduced the concept of orthogonality in the ordinary sense and of orthogonality in energy. They are both particular cases of the general concept of the orthogonality of the elements of a Hilbert space, the concept to which we devote the present section.

Two elements of a Hilbert space are said to be orthogonal if their scalar product equals zero. A finite or infinite sequence of elements which are mutually orthogonal is called an orthogonal system. If all the elements of a system are normalized, i.e. the norm of each element equals unity, then the system is called orthonormal. An orthonormal system $\{\varphi_n\}$ satisfies the equations

$$(\varphi_n, \varphi_m) = \begin{cases} 1, & m = n, \\ 0, & m \neq n. \end{cases}$$

The assertion about the linear independence of orthonorma elements holds true for an arbitrary Hilbert space. The theo rem of Pythagoras also holds true. If $\varphi = \varphi_1 + \varphi_2 + \ldots + \varphi$ and the elements are mutually orthogonal then

$$\|\varphi\|^2 = \sum_{k=1}^{n} \|\varphi_k\|^2. \tag{1}$$

The process of orthogonalization of linearly independent functions is extendable without change to the general Hilbert space.

In § 10 we introduced a definition of the completeness of a system of functions with respect to a certain class of func-

tions. In the general theory of Hilbert spaces another definition is widely used.

Let H be a Hilbert space and let $\{\varphi_n\}$ be an orthonormal system therein. We shall call this system *complete* in H if there exists no element H, apart from the identically zero element, which is orthogonal to all the elements of the system; otherwise the system is called incomplete.

If the system $\{\varphi_n\}$ is complete, then from the equations

$$(\varphi_n, \omega)=0, \quad n=1, 2, \ldots$$

it follows that $\omega=0$. This follows immediately from the definition of a complete system.

As in § 10, orthonormal systems are used to construct orthogonal series in an arbitrary Hilbert space.

We shall briefly dwell on this, bearing in mind that in general the elements of the space can befunctions which also assume complex values.

Let $\{\varphi_n\}$ be an orthonormal system in the Hilbert space H, and let φ be an arbitrary element of H. We pose the problem of finding the coefficients α_k such that the norm $\|\varphi-s_n\|$, where $s_n=\alpha_1\varphi_1+\alpha_2\varphi_2+\ldots+\alpha_n\varphi_n$, be a minimum.
We have

$$\|\varphi-s_n\|^2=(\varphi-s_n, \varphi-s_n)=\left(\varphi-\sum_{k=1}^{n}\alpha_k\varphi_k,\ \varphi-\sum_{m=1}^{n}\alpha_m\varphi_m\right)$$

$$=(\varphi, \varphi)-\sum_{k=1}^{n}\alpha_k(\varphi_k, \varphi)-\sum_{k=1}^{n}\overline{\alpha_k}(\varphi, \varphi_k)+\sum_{k=1}^{n}\alpha_k\overline{\alpha_k}.$$

We shall use the notation

$$(\varphi, \varphi_k)=a_k$$

$$\|\varphi-s_n\|^2=\|\varphi\|^2+\sum_{k=1}^{n}(\alpha_k\overline{\alpha_k}-\alpha_k\overline{a_k}-\overline{\alpha_k}a_k).$$

The numbers a_k are called the *Fourier coefficients* of the element φ with respect to the orthonormal system $\{\varphi_k\}$; now we add and subtract the quantity $a_k\overline{a_k}=|a_k|^2$ in the summation.

We then obtain

$$\|\varphi - s_n\|^2 = \|\varphi\|^2 + \sum_{k=1}^{n} (\alpha_k \bar{\alpha}_k - \alpha_k \bar{a}_k - \bar{\alpha}_k a_k + a_k \bar{a}_k) - \sum_{k=1}^{n} |a_k|^2$$

$$= \|\varphi\|^2 + \sum_{k=1}^{n} (\alpha_k - a_k)(\bar{\alpha}_k - \bar{a}_k) - \sum_{k=1}^{n} |a_k|^2,$$

or finally,

$$\|\varphi - s_n\|^2 = \|\varphi\|^2 + \sum_{k=1}^{n} |\alpha_k - a_k|^2 - \sum_{k=1}^{n} |a_k|^2.$$

Obviously the quantity $\|\varphi - S_n\|^2$ assumes its least value for $\alpha_k = a_k$ and

$$s_n = \sum_{k=1}^{n} a_k \varphi_k = \sum_{k=1}^{n} (\varphi, \varphi_k) \varphi_k,$$

$$\|\varphi - s_n\|^2 = \|\varphi\|^2 - \sum_{k=1}^{n} |a_k|^2. \tag{2}$$

The left-hand side of equation (2) is non-negative; hence there immediately follows the inequality

$$\sum_{k=1}^{n} |a_k|^2 \leq \|\varphi\|^2. \tag{3}$$

From this inequality, in turn, follows the convergence of the series

$$\sum_{k=1}^{\infty} |a_k|^2;$$

whence obviously,

$$\sum_{k=1}^{\infty} |a_k|^2 \leq \|\varphi\|^2. \tag{4}$$

The last inequality is called *Bessel's inequality.*

We consider the series

$$\sum_{k=1}^{\infty} a_k \varphi_k = \sum_{k=1}^{\infty} (\varphi, \varphi_k) \varphi_k,$$

known as the *Fourier series*, or *orthogonal series*, of the element φ with respect to the orthonormal system $\{\varphi_n\}$.

We shall prove, first of all, that this series converges,

provided that the space H is complete. We estimate the quantity

$$\|s_n - s_m\|^2 = \|\sum_{k=m+1}^{n} a_k \varphi_k\|^2.$$

By formula (1), because the φ_k are orthogonal and normal

$$\|s_n - s_m\|^2 = \sum_{k=m+1}^{n} |a_k|^2.$$

By virtue of the convergence of series (4), the quantity on the right tends to zero as $n \to \infty$, $m \to \infty$; by the definition of a complete space, there exists the limit

$$s = \lim_{n\to\infty} s_n = \sum_{k=1}^{\infty} a_k \varphi_k.$$

Further, the difference $\varphi - s$ is orthogonal to all φ_k, $k = 1, 2, \ldots$. In fact

$$(\varphi - s, \varphi_k) = (\varphi, \varphi_k) - (s, \varphi_k) = a_k - (s, \varphi_k)$$
$$= a_k - (s_n, \varphi_k) - (s - s_n, \varphi_k).$$

We take $n > k$. Since the system $\{\varphi_k\}$ is orthonormal, as is easily seen, $(s_n, \varphi_k) = a_k$ and, consequently,

$$(\varphi - s, \varphi_k) = (s_n - s, \varphi_k).$$

By Cauchy's formula

$$|(\varphi - s, \varphi_k)| \leq \|s_n - s\| \, \|\varphi_k\| = \|s_n - s\| \underset{n\to\infty}{\longrightarrow} 0,$$

and since $(\varphi - s, \varphi_k)$ does not depend upon n,

$$(\varphi - s, \varphi_k) = 0, \qquad k = 1, 2, \ldots$$

which proves our assertion.

We now suppose that the orthonormal system $\{\varphi_n\}$ is complete. In this case the element $\varphi - s$, which is orthogonal to all the elements of the system, equals zero and, consequently

$$\varphi = s = \sum_{k=1}^{\infty} a_k \varphi_k. \tag{5}$$

We thus arrive at the following theorem:

THEOREM 1. *If the space is complete, then the Fourier series*

in terms of any complete orthonormal system of any element of the space converges to this element.

By virtue of equation (1)

$$\|s_n\|^2=\sum_{k=1}^{n}|a_k|^2.$$

Further, by corollary 2 of theorem 1, § 43,

$$\|s\|^2=\lim_{n\to\infty}\|s_n\|^2=\sum_{k=1}^{\infty}|a_k|^2.$$

If the orthonormal system is complete, then $s=\varphi$ and we obtain the so called Parseval *equation*.

$$\|\varphi\|^2=\sum_{k=1}^{\infty}|a_k|^2. \tag{6}$$

It was established above that the sum of the squares of the Fourier coefficients of any element of the Hilbert space is convergent. The converse assertion is also valid:

THEOREM 2. *Let the constants a_k be such that the series $\sum_{k=1}^{\infty}|a_k|^2$ converges, and let $\{\varphi_k\}$ be an orthonormal system. Then the series*

$$\sum_{k=1}^{\infty}a_k\varphi_k \tag{7}$$

converges (in the sense of the convergence in the Hilbert space considered); the constants a_k are the Fourier coefficients of this sum, and for this sum the Parseval equation holds.

Denoting the nth partial sum of series (7) by s_n, we have

$$\|s_n-s_m\|^2=\left\|\sum_{k=m+1}^{n}a_k\varphi_k\right\|^2=\sum_{k=m+1}^{n}|a_k|^2\xrightarrow[m,\,n\to\infty]{}0,$$

whence follows the convergence of series (7). We denote its sum by s. We now find the Fourier coefficients of the element s:

$$(s,\ \varphi_k)=(s_n,\ \varphi_k)+(s-s_n,\ \varphi_k).$$

If $n>k$, then $(s_n,\ \varphi_k)=a_k$. Further,

$$|(s-s_n,\ \varphi_k)|\le\|s-s_n\|\,\|\varphi_k\|=\|s-s_n\|\xrightarrow[n\to\infty]{}0.$$

Hence it follows that $(s, \varphi_k) = a_k$. Finally

$$\|s\|^2 = \lim_{n\to\infty} \|s_n\|^2 = \lim_{n\to\infty} \sum_{k=1}^{n} |a_k|^2 = \sum_{k=1}^{\infty} |a_k|^2.$$

We now present a lemma which we shall often use in the sequel.

LEMMA. *An element orthogonal to all the elements of a dense set is equal to zero.*

Let H' be a dense set and let $(\varphi, \psi) = 0$ if $\psi \in H'$. We take an arbitrary element ω of our space. By the definition of a dense set, $\omega = \lim \psi_n$, where $\psi_n \in H'$. On the basis of corollary 1 § 43, $(\varphi, \omega) = \lim (\varphi, \psi_n) = 0$, so that $(\varphi, \psi_n) = 0$. Thus φ is orthogonal to any element of the space and, in particular, to itself. This gives $(\varphi, \varphi) = 0$. By axiom D, $\varphi = 0$.

We shall further dwell on a concept which has remained unconsidered in the previous sections and which nonetheless is of importance in the theory of the energy method. Particularly important are Hilbert spaces associated with a dense enumerable set. Such spaces are called *separable.*

The space $L_2(\Omega)$ (the space of functions that are quadratically summable in the domain Ω), which is particularly important in applications, is separable.

The proof of this assertion is based upon the extremely important theorem (see e. g. V. I. Smirnov [3] or I. P. Natanson [1]) by virtue of which a set of functions which are continuous in the closed domain $\overline{\Omega}$ is dense in $L_2(\Omega,)$ in other words, that any function whose square is summable in Ω can, with any degree of precision, be approximated in the mean by a continuous function.

We present the remainder of the argument assuming that the domain Ω is finite, only this case being of interest to us.

Let $u(P) \in L_2(\Omega)$. We enclose Ω within some parallelepiped Π and define the function $u(P)$ in $\Pi - \Omega$ by putting it equal to zero there. Obviously, after such an extrapolation $u(P) \in L_2(\Pi)$; by the above mentioned theorem $u(P)$ is approximated in the mean by some function which is con-

tinuous in Π. This function is approximated in Π uniformly (and *a fortiori* in the mean!) by a polynomial. This in turn, is approximated, as is completely obvious, by a polynomial with rational coefficients. This follows from Weierstrass's theorem [cf. 2, **154**]; the proof given there for functions in one variable is readily extended to the case of functions in many variables which are continuous in the parallelepiped.

Let $R(P)$ be this polynomial. Then the quantity

$$\int_{\Pi} |u(P) - R(P)|^2 d\Omega$$

is as small as required; *a fortiori* the smaller quantity

$$\int_{\Omega} |u(P) - R(P)|^2 d\Omega,$$

is as small as required, and this means that any function which is quadratically summable in the domain Ω is approximated in the mean in this domain by a polynomial with rational coefficients. But then the set of such polynomials is dense in $L_2(\Omega)$; at the same time it is known that this set is enumerable.

The importance of the property of separability is connected with the following theorem:

THEOREM 3. *In order that in a Hilbert space there exists a complete enumerable or finite orthonormal system, it is necessary and sufficient that this space be separable.*

Necessity: Let there exist in the space a complete finite or enumerable orthonormal sequence $\{\varphi_n\}$. By theorem 1 any element of the given space resolves into an orthogonal series

$$\varphi = \sum_{n=1}^{\infty} a_n \varphi_n, \qquad a_n = (\varphi, \varphi_n),$$

which converges in the sense of the metric of the given space. This implies, *inter alia*, that for a given $\varepsilon > 0$ it is possible to find an N such that

$$\left\| \varphi - \sum_{n=1}^{N} a_n \varphi_n \right\| < \frac{\varepsilon}{2}.$$

Further it is easy to select rational numbers a'_n, $n=1, 2, \ldots, N$ such that the inequality

$$\left\| \sum_{n=1}^{N} a_n \varphi_n - \sum_{n=1}^{N} a'_n \varphi_n \right\| = \sqrt{\left(\sum_{n=1}^{N} |a - a'_n|^2 \right)} < \frac{\varepsilon}{2}.$$

is satisfied. Then by the triangle inequality

$$\left\| \varphi - \sum_{n=1}^{N} a'_n \varphi_n \right\| < \varepsilon.$$

Thus, any element of the Hilbert space can be approximated with any degree of precision by a linear combination of a finite number of elements φ_n. This means that the set of such combinations is dense. But the coefficients of these combinations are also rational, and hence it follows that the set mentioned is enumerable (see, e.g. I. P. Natanson [1], Ch. I, § 3).

Sufficiency: Let the given Hilbert space be separable. Then it contains an enumerable dense set. The fact that this set is enumerable means that its elements can be numbered. Let these elements be

$$\psi_1, \psi_2, \ldots, \psi_n, \ldots. \tag{8}$$

Sequence (8) is such that, for no matter what element $\varphi \in H$ and for no matter what number $\varepsilon > 0$, there is an element ψ_n entering into (8) such that $\|\varphi - \psi_n\| < \varepsilon$. We subject the elements of (8) to an unusual "sifting": we strike from (8) all elements which are linearly dependent upon the previous ones. As a result we obtain a new sequence which we denote by

$$\omega_1, \omega_2, \ldots, \omega_n, \ldots; \tag{9}$$

its elements, taken any finite number at a time, are linearly independent. At the same time any of the elements of (8) can be linearly expressed by a finite number of the elements of (9). In fact, if ψ_n is an unrejected element then it is simply equal to some ω_n, if the element ψ_n has been rejected then,

by the construction, it depends linearly upon the previous ω_m. The fact that the sequence (8) is complete can now be formulated thus: for no matter what element $\varphi \in H$ and no matter what number $\varepsilon > 0$ it is possible to find a natural number n_0 and the coefficients $\alpha_1, \alpha_2, \ldots, \alpha_{n_0}$ such that

$$\left\| \varphi - \sum_{k=1}^{N} \alpha_k \omega_k \right\| < \varepsilon. \tag{10}$$

We orthogonalize sequence (9). Let $\{\varphi_n\}$ be the orthogonal system thus obtained. As we already know, each of the elements ω_k is expressed linearly in terms of $\varphi_1, \varphi_2, \ldots, \varphi_k$. Let

$$\omega_k = \sum_{j=1}^{k} b_{jk} \varphi_j.$$

Then

$$\sum_{k=1}^{n_0} \alpha_k \omega_k = \sum_{j=1}^{n_0} \beta_j \varphi_j; \qquad \beta_j = \sum_{k=j}^{n_0} b_{jk} \alpha_k,$$

and incquality (10) assumes the form

$$\left\| \varphi - \sum_{j=1}^{n_0} \beta_j \varphi_j \right\| < \varepsilon.$$

The last inequality proves that the orthonormal system $\{\varphi_n\}$ is complete in the given Hilbert space. We note that sequence (9) can be finite, and the system $\{\varphi_n\}$ will then also be finite.

§ 45. General definitions of functional and operator

From the general point of view which we are expounding in the present chapter the definitions given in § 4 of operator and functional are essentially connected with the space $L_2(\Omega)$.

In the present section we shall extend this definition for the case of an arbitrary Hilbert space.

We shall say that the functional $l\varphi$ is defined for a set M belonging to the Hilbert space H if each element $\varphi \in M$ of

the space generates some number $l\varphi$. The simplest example of a functional is the norm of an element. The scalar product (φ, ψ), in which the second factor ψ is fixed, is also a functional.

A functional is called linear if M is a linear set and

$$l(a_1\varphi_1 + a_2\varphi_2) = a_1 l\varphi_1 + a_2 l\varphi_2, \tag{1}$$

and bounded, if

$$|l\varphi| \leq N\|\varphi\|. \tag{2}$$

Here, a_1, a_2, N are constants, φ_1, φ_2, φ, are elements of the Hilbert space. The scalar product (φ, ψ) where ψ is fixed, provides an example of a linear bounded set.

Its linearity follows from axiom B § 42, and its boundedness, from Cauchy's formula which shows that it is possible to put $N = \|\psi\|$. The norm of an element is a bounded functional (here $N = 1$) but it is not linear.

If l is a linear functional, then obviously $l0 = 0$. The least of the numbers N satisfying inequality (2) is called the *norm* of the bounded functional $l\varphi$ and is denoted by $\|l\|$. Thus

$$|l\varphi| \leq \|l\|\,\|\varphi\|, \tag{3}$$

where it is not possible to replace $\|l\|$ by a smaller number.

The functional $l\varphi$ is called continuous if

$$\lim_{\varphi \to \psi} l\varphi = l\psi. \tag{4}$$

The norm of an element is a continuous functional. This follows from theorem 1 § 43.

Every linear bounded functional which is defined for the whole space is continuous. In fact, let $\varphi \to \psi$. Using the linearity of the functional and inequality (2), we find

$$|l\varphi - l\psi| = |l(\varphi - \psi)| \leq N\|\varphi - \psi\| \underset{\varphi \to \psi}{\to} 0. \tag{5}$$

From what has been proved there follows, *inter alia*, the continuity of scalar product. The converse assertion can also be proved: any linear and continuous functional is bounded.

THEOREM 1. *Every bounded functional linear in H has the form of the scalar product*

$$l\varphi=(\varphi,\ \psi), \tag{6}$$

where ψ is a fixed element of the space H. The element ψ is determined uniquely.

We shall carry out the proof, assuming that H is separable. By theorem 3 § 44 there exists in H a complete, enumerable, orthonormal system. Let $\{\varphi_n\}$ be such a system. We shall use the notation $a_n=\overline{l\varphi_n}$ and we shall prove that the series $\sum_{n=1}^{\infty}|a_n|^2$ converges. We put

$$\psi_n=\sum_{k=1}^{n}a_k\varphi_k;$$

then

$$l\psi_n=l(a_1\varphi_1+\ \dots\ +a_n\varphi_n)=a_1l\varphi_1+\ \dots\ +a_nl\varphi_n=\sum_{k=1}^{n}|a_k|^2.$$

Since the functional $l\varphi$ is bounded, $|l\varphi_n|\leq\|l\|\,\|\psi_n\|$, or

$$\sum_{k=1}^{n}|a_k|^2\leq\|l\|\sqrt{\left(\sum_{k=1}^{n}|a_k|^2\right)}.$$

Dividing by $\sqrt{\left(\sum_{k=1}^{n}|a_k|^2\right)}$ and squaring, we obtain

$$\sum_{k=1}^{n}|a_k|^2\leq\|l\|^2.$$

Thus, the convergence of our series follows and

$$\sum_{k=1}^{\infty}|a_k|^2\leq\|l\|^2. \tag{7}$$

The series

$$\psi=\sum_{k=1}^{\infty}a_k\varphi_k=\lim_{n\to\infty}\psi_n \tag{*}$$

converges.

Now let φ be an arbitrary element of the space. We resolve it into a Fourier series:

$$\varphi=\sum_{k=1}^{\infty}\alpha_k\varphi_k, \quad \alpha_k=(\varphi, \varphi_k)$$

and let $\varphi^{(n)}=\sum_{k=1}^{n}\alpha_k\varphi_k$, so that $\varphi=\lim_{n\to\infty}\varphi^{(n)}$. We have

$$l\varphi^{(n)}=\sum_{k=1}^{n}\alpha_k l\varphi_k=\sum_{k=1}^{n}(\varphi, \varphi_k)\overline{a}_k.$$

We place the numerical factor $\overline{a}_k$ within the scalar product and attach it to the second factor. Then

$$l\varphi^{(n)}=\sum_{k=1}^{n}(\varphi, a_k\varphi_k)=\left(\varphi, \sum_{k=1}^{n}a_k\varphi_k\right)=(\varphi, \psi_n).$$

Now let $n\to\infty$. Using the continuity both of functional $l\varphi$ and of scalar product we find that

$$l\varphi=(\varphi, \psi),$$

which was to be proved. From Cauchy's formula

$$|l\varphi|=|(\varphi, \psi)|\leq\|\psi\|\,\|\varphi\|$$

it follows that $\|l\|\leq\|\psi\|$; on the other hand, from (7) it follows that $\|\psi\|\leq\|l\|$. Comparing both inequalities, we conclude that

$$\|l\|=\|\psi\|. \tag{8}$$

We now prove that the element ψ in formula (6) is uniquely defined. Let us suppose the opposite and let $l\varphi=(\varphi, \psi')$, where $\psi'\neq\psi$. Subtracting the last equation from (6), we obtain

$$(\varphi, \psi'-\psi)=0.$$

Thus $\psi'-\psi$ is orthogonal to any element of the space In particular we can put $\varphi=\psi'-\psi$. Then $(\psi'-\psi, \psi'-\psi)=0$ By axiom D we have $\psi'-\psi=0$, or $\psi'=\psi$.

It is sometimes necessary to consider functionals which are defined not for the whole space but for some linear set which constitutes a part of the space.

In this case there often arises the problem of the extension of the functional, i.e. of the definition of the functional for a wider set. The following is essential in this respect.

THEOREM 2. *A linear bounded functional specified for a dense linear set can be extended for the entire space without altering the norm of the functional.*

Let H' be a linear set dense in the Hilbert space H, and let a bounded linear functional $l\varphi$ be specified in H'. We take an element $\varphi \in H$. Since H' is dense in H, it is possible to find a sequence $\varphi_n \in H'$, such that $\varphi = \lim \varphi_n$. We have

$$|l\varphi_m - l\varphi_n| = |l(\varphi_m - \varphi_n)| \leq \|l\| \, \|\varphi_m - \varphi_n\| \xrightarrow[m,\, n\to\infty]{} 0,$$

since the sequence $\{\varphi_n\}$ is convergent. By virtue of Cauchy's test for the existence of a limit of a numerical sequence, the limit

$$\lim_{n\to\infty} l\varphi_n$$

exists.

By defining

$$l\varphi = \lim_{n\to\infty} l\varphi_n,$$

we extend the functional $l\varphi$ for all of the space H.

It is easily seen that the extended functional remains linear and that its norm is unchanged. The latter is clear from the fact that

$$|l\varphi| = \lim_{n\to\infty} |l\varphi_n| \leq \lim_{n\to\infty} \|l\| \, \|\varphi_n\| = \|l\| \, \|\varphi\|.$$

It can be proved that the extension indicated in the theorem is unique.

We shall say that *operator* A is defined for some set D, if each element $\varphi \in D$ generates according to some law one and only one element of Hilbert space. We shall write $\psi = A\varphi$. The set D is called *the field of definition* of the operator $A\varphi$, and the set D' of all possible elements ψ, *the field of values* of the operator.

It can be said that $A\varphi$ is a function which transforms an element $\varphi \in D$ into an element $\psi \in D'$. We shall denote the

field of definition of the operator A, as earlier, by D_A. It should be noted that the definition of operator necessarily involves its field of definition. Two operators A and B are considered equal if their fields of definition coincide and if for each element φ which enters their field of definition $A\varphi = B\varphi$. If D_A is part of D_B and for each element $\varphi \in D_A$ the equation $A\varphi = B\varphi$ is true, then operator B is called an *extension* of operator A.

An operator A is called *linear* if its field of definition is a linear set and (a_1, a_2 are constants)

$$A(a_1\varphi_1 + a_2\varphi_2) = a_1 A\varphi_1 + a_2 A\varphi_2. \tag{9}$$

If A is a linear operator, then $A0 = 0$. An operator is called *continuous* if

$$\lim_{\varphi \to \psi} A\varphi = A\psi, \tag{10}$$

and *bounded* if

$$\|A\varphi\| \leq C\|\varphi\|, \qquad C = \text{const.} \tag{11}$$

The least of the constants C satisfying inequality (11) is known as the *norm* of the bounded operator A and is denoted by the symbol $\|A\|$. Obviously

$$\|A\varphi\| \leq \|A\|\|\varphi\|. \tag{12}$$

A linear bounded operator is continuous — this is proved in the same way as for functionals. The converse assertion is also valid: a linear continuous operator is bounded.

In what follows we shall consider only linear operators; in speaking of an operator we shall always assume that it is linear.

A theorem analogous to theorem 2 for functionals is also valid for operators:

THEOREM 3. *A linear bounded operator specified for a dense set can be extended for the entire space without altering the norm of the function.*

The proof of this theorem coincides almost word for word with the proof of theorem 2. Let the field of definition D_A

of operator A be dense in H. If $\varphi \in H$ then $\varphi = \lim \varphi_n$, $\varphi_n \in D_A$. We have

$$\|A\varphi_n - A\varphi_m\| = \|A(\varphi_n - \varphi_m)\| \leq \|A\| \, \|\varphi_n - \varphi_m\| \xrightarrow[n, m \to \infty]{} 0.$$

By virtue of the completeness of the space, the limit exists.

$$\lim_{n \to \infty} A\varphi_n.$$

Putting

$$A\varphi = \lim_{n \to \infty} A\varphi_n,$$

we extend the operator for the entire space. Precisely as in theorem 2, it can be proved that $\|A\|$ then remains unchanged and that such an extension of a bounded operator is unique.

The operator which transforms each element of the space into itself is called the *identity operator*. We shall denote it by E, so that $E\varphi \equiv \varphi$. The operator which transforms each element of the space into the null element is called the null operator. Both these operators are linear and bounded; the norm of the identity operator equals unity, the norm of the null operator is zero.

The operators $A+B$ and AB defined by the relations

$$(A+B)\varphi = A\varphi + B\varphi, \qquad AB\varphi = A(B\varphi)$$

are called the sum and product of the operators A and B.

In general $AB \neq BA$; thus, for example, if

$$A\varphi = \frac{\mathrm{d}\varphi}{\mathrm{d}x}; \quad B\varphi = x \int_a^b \varphi(s)\mathrm{d}s,$$

then

$$AB\varphi = \int_a^b \varphi(s)\mathrm{d}s,$$

$$BA\varphi = x \int_a^b \varphi'(s)\mathrm{d}s = x[\varphi(b) - \varphi(a)] \neq AB\varphi.$$

The following notation is commonly used: $AA = A^2$, $AA^2 = A^3$, etc., in general $A^n = AA^{n-1}$. Obviously $A^m A^n = A^{m+n}$.

Let the operators A and B be bounded. By the triangle inequality

$$\|(A+B)\varphi\| \leq \|A\varphi\| + \|B\varphi\|,$$

and, since A and B are bounded,

$$\|(A+B)\varphi\| \leq (\|A\| + \|B\|)\|\varphi\|.$$

Hence from the definition of the norm of an operator

$$\|A+B\| \leq \|A\| + \|B\|. \tag{13}$$

Similarly

$$\|AB\varphi\| = \|A(B\varphi)\| \leq \|A\|\,\|B\varphi\| \leq \|A\|\,\|B\|\,\|\varphi\|$$

and, consequently,

$$\|AB\| \leq \|A\|\,\|B\|. \tag{14}$$

Of extreme importance in all that follows is the concept of *inverse operator.*

Let D_A be the field of definition of operator A and let R_A be its field of values. By definition, to each element of D_A there corresponds one and only one element of R_A. On the other hand, to each element of R_A there corresponds at least one element of D_A. We suppose that to each element of R_A there corresponds only one element of D_A; this correspondence defines some operator B having R_A as its field of definition and D_A as its field of values. Operator B is called the inverse of the operator A. It is also obvious that the operator A is the inverse of B. We shall write $B=A^{-1}$; instead of $(A^{-1})^n$ we write A^{-n}.

From the definition of inverse operator there follows the following: if A and B are mutually inverse operators and $\varphi \in D_A$, then $BA\varphi=\varphi$, similarly if $\psi \in D_B$ then $AB\psi=\psi$. It is obvious that the inverse of a linear operator is also linear.

The following is almost obvious:

THEOREM 4. *In order that operator A have an inverse, it is necessary and sufficient that the equation*

$$A\varphi=0 \tag{15}$$

have a unique solution $\varphi=0$.

Let the inverse operator B exist, and let φ_0 be the solution of equation (15). Then, in every case, $\varphi_0 \in D_A$ as was stated above, $BA\varphi_0 = \varphi_0$. But $BA\varphi_0 = B0 = 0$ and, consequently, $\varphi_0 = 0$. We now suppose that the condition of the theorem is satisfied. Let ψ be an arbitrary element from the field of values of operator A. We shall prove that only one element from the field of definition of A corresponds to it — this will establish the existence of the operator A^{-1}. Let us suppose the opposite to be true: let two elements φ_1 and φ_2 from the field of definition of the operator A correspond to the element ψ. This means that $A\varphi_1 = \psi$ and $A\varphi_2 = \psi$. But then $A(\varphi_1 - \varphi_2) = 0$; by the condition $\varphi_1 - \varphi_2 = 0$ or $\varphi_1 = \varphi_2$, contrary to assumption.

THEOREM 5. *In order that the operator A^{-1} be bounded, it is necessary and sufficient that a constant $k > 0$ exists such that for all $\varphi \in D_A$*

$$\|A\varphi\| \geq k\|\varphi\|. \tag{16}$$

Let the inverse operator $B = A^{-1}$ be bounded. Then

$$\|B\psi\| \leq \|B\|\|\psi\|.$$

We put $\varphi = B\psi$. Then $\psi = A\varphi$ and the last inequality transforms to inequality (16), if we put $k = \dfrac{1}{\|B\|}$. Conversely, let inequality (16) be valid. Let φ_1, satisfy the equation $A\varphi_1 = 0$. Then

$$\|\varphi_1\| \leq \frac{1}{k}\|A\varphi_1\| = 0,$$

whence $\varphi_1 = 0$. By theorem 4 the inverse operator $B = A^{-1}$ exists. We put $A\varphi = \psi$. Then $\varphi = B\psi$ and from (16) it follows that

$$\|B\psi\| \leq \frac{1}{k}\|\psi\|.$$

The last inequality shows that B is a bounded operator, and its norm does not exceed $\dfrac{1}{k}$.

Operator A is called *symmetric* if it is defined for a dense set and if for any elements φ and ψ from the field definition of this operator the identity

$$(A\varphi, \psi) = (\varphi, A\psi) \tag{17}$$

is valid.

The following theorem gives a simple criterion for the symmetry of an operator in a complex Hilbert space; in a real space this criterion is obviously invalid.

THEOREM 6. *In order that an operator A, defined for a dense set, be symmetric, it is necessary and sufficient that the scalar product $(A\varphi, \varphi)$ be real.*

Necessity. Let the operator A be symmetric. Putting $\psi = \varphi$ in (3) we obtain $(A\varphi, \varphi) = (\varphi, A\varphi)$ or, by axiom B § 42, $(A\varphi, \varphi) = \overline{(A\varphi, \varphi)}$. As it equals its conjugate, the quantity $(A\varphi, \varphi)$ is real.

Sufficiency. The identity

$$\begin{aligned} 4A(\varphi, \psi) = {} & (A(\varphi+\psi), \varphi+\psi) - (A(\varphi-\psi), \varphi-\psi) \\ & + i[(A(\varphi+i\psi), \varphi+i\psi) - (A(\varphi-i\psi), \varphi-i\psi)] \end{aligned}$$

is readily verified. By hypothesis all scalar products o the right are real. Interchanging φ and ψ, we find

$$\begin{aligned} 4(A\psi, \varphi) = {} & (A(\psi+\varphi), \psi+\varphi) - (A(\psi-\varphi), \psi-\varphi) \\ & + i[(A(\psi+i\varphi), \psi+i\varphi) - (A(\psi-i\varphi), \psi-i\varphi)]. \end{aligned}$$

From the fundamental properties of scalar product it follows that

$$\begin{aligned} (A(\psi-\varphi), \psi-\varphi) &= (A(\varphi-\psi), \varphi-\psi) \\ (A(\psi+i\varphi), \psi+i\varphi) &= (iA(\varphi-i\psi), i(\varphi-i\psi)) \\ &= (A(\varphi-i\psi), \varphi-i\psi), \\ (A(\psi-i\varphi), \psi-i\varphi) &= (-iA(\varphi+i\psi), -i(\varphi+i\psi)) \\ &= (A(\varphi+i\psi), \varphi+i\psi). \end{aligned}$$

Hence it is easily seen that $(A\varphi, \psi) = \overline{(A\psi, \varphi)}$ or, by axiom A § 42, $(A\varphi, \psi) = (\varphi, A\psi)$, i.e. that operator A is symmetric.

Essentially we retain the definitions given in § 6 of a positive definite and of a positive-bounded-below operator. Thus if we are concerned with a complex Hilbert space, then the operator Au, defined for a set which is dense in the Hilbert space, is called positive definite if for any non-zero element from the field of definition of the operator the inequality $(Au, u) > 0$ is valid; the operator Au is called *positive-bounded-below* if for any element from the field of definition of the operator the much stronger inequality

$$(Au, u) \geq \gamma^2 \|u\|^2,$$

is satisfied, where γ^2 is a positive constant.

In the case of a complex Hilbert space there is no need to stipulate the symmetry of an operator which is positive definite or positive-bounded-below — it follows from theorem 5. In the case of a real space it is necessary to retain the stipulation of symmetry.

For an operator A which is positive definite in an arbitrary Hilbert space theorem 1 § 11 about the uniqueness of the solution of the equation $Au = f$ still holds.

§ 46. The general formulation of the variational problem and its solution

We shall consider the equation

$$Au = f, \tag{1}$$

in which u is a sought and f a given element of the Hilbert space and A is a positive definite or a positive-bounded-below operator.

THEOREM 1. *Let A be a positive definite operator. If equation* (1) *has a solution then this solution minimizes the functional*

$$F(u) = (Au, u) - (u, f) - (f, u). \tag{2}$$

Conversely the element of Hilbert space which realizes the minimum of functional (2) *satisfies equation* (1).

Firstly we note that the fields of definition of the operator Au and of the functional $F(u)$ coincide. This immediately follows from the form of expression (2). Further $F(u) = (Au, u) - 2\mathrm{Re}(u,f)$ and, since A is a positive definite operator, $F(u)$ assumes only real values.

Now let u_0 satisfy equation (1):

$$Au_0 - f = 0.$$

Further, let v be an arbitrary element from the field of definition D_A of operator A. We put $v - u_0 = \eta$, so that $v = u_0 + \eta$. We have

$$F(v) = (A(u_0+\eta), u_0+\eta) - (u_0+\eta, f) - (f, u_0+\eta).$$

Expanding the brackets and using the symmetry of A, we readily find that

$$F(v) = F(u_0) + (Au_0 - f, \eta) + (\eta, Au_0 - f) + (A\eta, \eta).$$

But $Au_0 = 0$, thus $F(v) = F(u_0) + (A\eta, \eta)$ and because A is a positive definite operator, $F(v) > F(u_0)$, i.e., the functional $F(u)$ assumes its minimum value at $u = u_0$.

Conversely, let u_0 minimize the functional $F(u)$. Let η be an arbitrary element from D_A. The set D_A as the field of definition of a linear operator is itself linear, therefore $\lambda\eta$, where λ is a constant number, also enters into D_A; and similarly $u_0 + \lambda\eta \in D_A$. By hypothesis

$$F(u_0 + \lambda\eta) \geq F(u_0).$$

We shall assume that the number λ is real. Using the symmetry of A, the last inequality is readily reduced to the form

$$2\lambda \mathrm{Re}[(Au_0 - f, \eta)] + \lambda^2 (A\eta, \eta) \geq 0.$$

This is only possible if

$$\mathrm{Re}[(Au_0 - f, \eta)] = 0.$$

Replacing η by $i\eta$, we obtain

$$\mathrm{Im}[(Au_0 - f, \eta)] = 0.$$

Hence it follows that $(Au_0-f, \eta)=0$. By virtue of the last equation, the element Au_0-f of the Hilbert space is orthogonal to all the elements of the dense set D_A. By the lemma of § 44, $Au_0-f=0$ i. e. u_0 satisfies equation (1).

NOTE: If the given Hilbert space is real, $(u, f)=(f, u)$ and the functional (2) assumes the simpler form

$$F(u)=(Au, u)-2(u, f),$$

which has been widely used in the previous chapters.

Theorem 1 just proved allows us to assert that the problem of the solution of equation (1) is equivalent to the problem of finding the minimum of functional (2).

The exact meaning of this assertion is as follows: let an operator A be defined for a linear set M which is dense in the given Hilbert space H and let it be positive definite for this linear set; formula (2) shows that the functional $F(u)$ is also defined for the same linear set. Now let there exist an element $u_0 \in M$ which satisfies equation (1); then $F(u) > F(u_0)$, where u is any element of M. Conversely, if $u_0 \in M$ and for any $u \in M$ the inequality $F(u) > F(u_0)$ is valid, then u_0 satisfies equation (1).

It is necessary to make two observations in connexion with what has been stated:

1. Our assertion assumes the existence of an element which satisfies equation (1) or minimizes functional (2), while it is important for us to prove the very existence of such an element and to find methods, approximate or exact, for finding it. It is impossible to do this by merely assuming that operator A is positive definite; *in what follows we shall assume that it is positive-bounded-below.*

2. It may happen that the operator is defined for an insufficiently wide linear set. In such a case the problem of the minimum of functional (2) can have no solution. It will exist however if the linear set M, and with it the operator, are sufficiently extended. Below we shall show that for a positive-bounded-below operator an extension of the linear set M,

for which the problem of the minimum of the functional $F(u)$ has a solution, is always possible.

(As far as we know, the general solution of this problem was first obtained by K. Friedrichs [2]. Our solution (see S. G. Mikhlin [4] and [8]) differs from Friedrichs's solution both in form and in method of proof.)

The necessity of an extension of the operator is readily explained as follows. Let S be a surface domain bounded by the line L. We consider the operator $\nabla^4 w$. In order to define it completely it is necessary to specify its field of definition, or in other words, we must characterize the set of functions for which this operator is specified. To this end we state the boundary conditions which these functions satisfy and the number of continuous derivatives that they possess. Let functions from the field of definition of our operator satisfy the boundary conditions

$$\omega|_L = 0, \qquad \left.\frac{\partial w}{\partial n}\right|_L = 0.$$

From the mechanical viewpoint we are concerned here with a plate with a rigidly fixed edge. Inasmuch as fourth derivatives enter into the operator ∇^4, it is natural to require that these derivatives be continuous. We shall do this and pose the problem of the bending of a plate under the action of normal loads which vary *discontinuously*. This reduces to the integration under the above mentioned boundary conditions of the equation

$$\nabla^4 \omega = \frac{\partial^4 w}{\partial x^4} + 2\frac{\partial^4 w}{\partial x^2 \partial y^2} + \frac{\partial^4 w}{\partial y^4} = \frac{q(x, y)}{D},$$

where $q(x, y)$ is a discontinuous function. But this equation cannot have a solution in the field of the operator ∇^4 defined above, since in the solution of our equation at least one of the fourth-order derivatives

$$\frac{\partial^4 w}{\partial x^4}, \quad \frac{\partial^4 w}{\partial x^2 \partial y^2}, \quad \frac{\partial^4 w}{\partial y^4}$$

is discontinuous. If we are going to find the solution of our equation as an element from the field of definition of the operator ∇^4 then, obviously it is initially necessary to extend this field in such a way that it will also contain functions which possess discontinuous fourth-order derivatives.

Having made these remarks, we pass on to the solution of the problem of the minimum of the functional (2). This solution will be constructed on the assumption that operator A is positive-bounded-below (the case where A is positive definite but not positive-bounded-below is considered in the author's paper [14], so that

$$(Au, u) \geq \gamma^2 \|u\|^2.$$

It is essential for the following to introduce a new Hilbert space which we referred to briefly in § 42 and which we denoted there by H_0 or by H_A.

For the linear set M we introduce the new scalar product (to distinguish it from the old one we shall denote it by square brackets) by putting as a definition

$$[u, v] = (Au, v). \tag{3}$$

We shall prove that such a definition is proper, i.e. that it satisfies axioms A–D § 42.

Axiom A. The operator A, being positive-bounded-below, is symmetric; therefore, if $u \in M$ and $v \in M$, then

$$[u, v] = (Au, v) = (u, Av) = \overline{(Av, u)} = \overline{[v, u]}.$$

Axiom B.

$$[a_1u_1 + a_2u_2, v] = (A(a_1u_1 + a_2u_2), v)$$
$$= a_1(Au_1, v) + a_2(Au_2, v) = a_1[u_1, v] + a_2[u_2, v].$$

Axioms C and D. $(Au, u) \geq \gamma^2 \|u\|^2 \geq 0$; $\varphi(Au, u) = 0$ then necessarily $\|u\| = 0$ and, consequently, $u = 0$.

By introducing a new scalar product we have transformed M into a Hilbert space. We shall denote the norm in this space by $\|u\|$, so that

$$\|u\|^2 = [u, u] = (Au, u), \quad u \in M. \tag{4}$$

Thus the inequality

$$\|u\| \leq \frac{1}{\gamma} \|u\| \tag{5}$$

follows.

Our new Hilbert space may be incomplete, in which case we supplement it in the usual manner by introducing limiting elements. We shall also denote this complete Hilbert space by H_0 or sometimes by H_A. For the quantities $[u, v]$ and $\|u\|$ we retain the names of energy product and energy norm introduced earlier. We further note that if $H=L_2(\Omega)$, then H_0 consists of the functions which we once called functions with finite energy.

In the cases where it is important to stress that the derived values are connected with operator A we shall, as before, write $[u, v]_A$, $\|u\|_A$ instead of $[u, v]$, $\|u\|$.

If n is an element of the space H_0, then by definition either $u \in M$ or a sequence $u_n \in M$ exists such that $\|u_n - u\| \xrightarrow[n\to\infty]{} 0$. Hence it follows that the linear set M is dense in H_0.

THEOREM 2. *All the elements of the space H_0 also belong to the space H.*

More precisely it is possible to place in correspondence with each element from H_0 one and only one element from H, and to different elements of H_0 there correspond different values of H.

Our assertion is obvious if $u \in M$ – it suffices to let the element u correspond to itself. It remains to consider the case where u is a limiting element of the space H_0. In this case, by definition, there exists a sequence $u_n \in M$ such that $\|u_n - u\| \xrightarrow[n\to\infty]{} 0$. From the triangle inequality

$$\|u_n - u_m\| \leq \|u_n - u\| + \|u_m - u\|$$

and consequently, $\|u_n - u_m\| \xrightarrow[n,m\to\infty]{} 0$. From inequality (5) which is at present established for elements of the linear set M, it follows that $\|u_n - u_m\| \xrightarrow[n,m\to\infty]{} 0$. This means that the sequence $\{u_n\}$ converges in the space H (assumed complete) to one of its elements u' so that $\|u' - u_n\| \xrightarrow[n\to\infty]{} 0$. Identifying u and u' we have established the first part of our assertion. We now prove its second part.

Let a given element from H correspond to both the elements $u^{(1)}$ and $u^{(2)}$ from H_0. Then the null element from H corresponds to the difference $u = u^{(1)} - u^{(2)}$. We shall prove that u is the null element from H_0. Let u be the limit (in the sense of convergence in H_0) of the sequence $u_n \in M$, so that $\|u_n - u\| \xrightarrow[n\to\infty]{} 0$. Since to element u in space H there corresponds the null

element, then by the correspondence law established above $\|u_n\| \to 0$. Hence it follows that $(f, u_n) \xrightarrow[n\to\infty]{} 0$ where f is any fixed element of H. In fact by Cauchy's inequality

$$|(f, u_n)| \leq \|f\| \, \|u_n\| \to 0.$$

We select an arbitrary element φ in the linear set M and put $f = A\varphi$. Then $(A\varphi, u_n) \to 0$. But $(A\varphi, u_n) = [\varphi, u_n]$, so that both the elements φ and u_n enter into M; hence $[\varphi, u_n] \to 0$. Proceeding to the limit we obtain

$$[\varphi, u] = 0.$$

Thus in the space H_0 the element u is orthogonal to all the elements of the linear set M which is dense in H_0. By the lemma of § 43 u is the null element of H_0. We note that inequality (5) is satisfied not only in M but also in the whole of the space H_0. In fact let u be an element of H_0 which does not enter into M. Then a sequence $u_n \in M$ exists such that $\|u_n - u\| \to 0$. As follows from the proof of theorem, in this case $\|u_n - u\| \to 0$. Hence it follows that $\|u_n\| \to \|u\|$ and $\|u_n\| \to \|u\|$. Since $u_n \in M$ inequality (5) is valid for u_n:

$$\|u_n\| \leq \frac{1}{\gamma} \|u_n\|.$$

Passing to the limit as $u \to \infty$, we obtain

$$\|u\| \leq \frac{1}{\gamma} \|u\|.$$

Now the problem of the minimum of the functional $F(u)$ is readily solved. Firstly $(Au, u) = \|u\|^2$. This equation extends the functional, which initially we only defined for M, to the whole of the space H_0. We shall now understand by u in $F(u)$ any element of the space H_0 and correspondingly we replace (Au, u) by $\|u\|^2 = [u, u]$. Further the linear functional (u, f) is bounded in H_0. In fact

$$|(u, f)| \leq \|f\| \|u\| \leq \frac{\|f\|}{\gamma} \|u\|.$$

Theorem 1 § 45 asserts the existence of a unique element $u_0 \in H_0$ such that, for any $u \in H_0$

$$(u, f) = [u, u_0]. \tag{6}$$

Now $(f, u) = [u_0, u]$ and

$$F(u) = [u, u] - [u, u_0] - [u_0, u] = [u - u_0, u - u_0] - [u_0, u_0]$$

or more concisely,

$$F(u) = \|u - u_0\|^2 - \|u_0\|^2. \tag{7}$$

It immediately follows from theorem (7) that $F(u)$ *becomes a minimum when* $u = u_0$.

Thus the problem of the minimum of functional (2) is solved. Theorem 1 § 45 mentioned above allows us to give an explicit expression for the element u_0 in the form of a series analogous to series 2 § 12. We shall assume that H_0 is separable [as the author has proved ([11] p. 33, theorem 2) it suffices for this that the initial space H be separable]. By theorem 3 § 44 there exists in H_0 a complete, enumerable orthonormal system, in other words, there exists a system of functions φ_n, $n = 1, 2, \ldots$ which is orthonormal and complete in energy.

In our case the functional lu referred to in theorem 1 § 45 is equivalent to (u, f) and the element ψ to u_0. By formula (*) § 45,

$$u_0 = \sum_{k=1}^{\infty} a_k \varphi_k,$$

where $a_k = \overline{l\varphi_n} = \overline{(\varphi_k, f)} = (f, \varphi_k)$. Finally

$$u_0 = \sum_{k=1}^{\infty} (f, \varphi_k)\varphi_k. \tag{8}$$

We further note that u_0 may not belong to M, so that the equation $Au = f$ may not have a solution if the operator A is only defined for M. This is connected with what we have said above about the necessity of extending this operator.

We now study the solution of the minimum problem that we have obtained more carefully. Formula (6) generates from each element f of H a certain unique element $u=u_0 \in H_0$ which realizes the minimum of functional (2). Thus formula (6) defines the operator $u_0=Gf$, whose field of definition is equivalent to the given Hilbert space, and whose field of values is a part of H_0. (In special cases the field of values Gf can coincide with H_0.) With the aid of operator Gf the formula (6) can be represented in the form

$$(u, f)=[u, Gf], \quad u \in H_0. \tag{9}$$

We shall prove that the operator G is bounded and positive definite. Firstly this operator is linear. In fact if $f=a_1f_1+a_2f_2$, then

$$\begin{aligned}(u, f)&=\bar{a}_1(u, f_1)+\bar{a}_2(u, f_2)=\bar{a}_1[u, Gf_1]+\bar{a}_2[u, Gf_2]\\&=[u, a_1Gf_1+a_2Gf_2].\end{aligned}$$

On the other hand $(u, f)=[u, Gf]$. Comparing the two expressions for (u, f) we find that

$$[u, Gf-a_1Gf_1-a_2Gf_2]=0; \qquad u \in H_0.$$

By the lemma of § 43, $Gf-a_1Gf_1-a_2Gf_2=0$, or

$$G(a_1f_1+a_2f_2)=a_1Gf_1+a_2Gf_2,$$

i.e. the operator G is linear. Further, by Cauchy's inequality,

$$|[u, Gf]| = |(u, f)| \leq \|f\|\|u\|$$

or using inequality (5),

$$|[u, Gf]| \leq \frac{\|f\|}{\gamma}\|u\|.$$

We now put $u=Gf$, giving

$$\|Gf\|^2 \leq \frac{\|f\|}{\gamma}\|Gf\|,$$

or

$$\|Gf\| \leq \frac{\|f\|}{\gamma}. \tag{*}$$

But, by inequality (5), $\|Gf\| \leq \frac{|Gf|}{\gamma}$. Substituting this in (*), we obtain:

$$\|Gf\| \leq \frac{\|f\|}{\gamma^2},$$

whence

$$\|G\| \leq \frac{1}{\gamma^2}, \tag{10}$$

i.e. operator G is bounded. Finally, by formula (9)

$$(Gf, f) = [Gf, Gf] = |Gf|^2 > 0.$$

This inequality shows that the operator Gf is positive definite.

NOTE: We shall consider that $f \in H_0$. Thus we consider Gf as an operator which is defined in the space H_0. It is easily seen that in this case it remains bounded. In fact, we replace $\|f\|$ by the larger quantity $\frac{|f|}{\gamma}$ in (*). Then $|Gf| \leq \frac{|f|}{\gamma^2}$ or $|G| \leq \frac{1}{\gamma^2}$. Further, by formula (9) $[Gf, f] = [f, Gf] = \overline{(f, f)} = (f, f) \geq 0$. Hence, as above, there follows the positive-definiteness of Gf, but this time in the space H_0.

If $Gf = 0$, then formula (9) gives $(u, f) = 0$. Thus the element f is orthogonal in H to all the elements of H_0 and, in particular, to all the elements of the original linear set M which is dense in H. By the lemma of § 43 we have $f = 0$. Thus the equation $Gf = 0$ has the unique solution $f = 0$; hence, by theorem 4 § 45 the inverse operator to G, G^{-1} exists.

THEOREM 3. *The operator G^{-1} is an extension of the operator A.* It suffices to prove that $G^{-1}u$ is equivalent to Au for M, i. e. that $G^{-1}u = Au$ if $u \in M$. Let u_0 be an arbitrary element from M.

We denote Au_0 by f. Then $Au_0 = f$. By theorem 1 u_0 minimizes functional (2). As proved in this section, $u_0 = Gf$. Hence $f = G^{-1}u_0$ and consequently $G^{-1}u_0 = Au_0$, which was to be proved.

The solution we have constructed of the problem of the minimum of functional (2) belongs to $D_A = M$. By theorem 1,

this solution is simultaneously the solution of equation (1). If it does not belong to M we shall, nonetheless consider it as the (generalized) solution of equation (1). Thus we extend the operator A. It can be proved that under such an extension operator A remains positive-bounded-below.

It would be easy to extend the Ritz method to the case of an operator which is specified in an arbitrary Hilbert space. We shall leave this to the reader and merely note that equations (8) and (8_1) § 14 retain their form in the general case. We further note that in the general case the theorems of § 31 about eigenvalues and eigenfunctions remain valid.

§ 47. The method of minimum surface integrals

In order to explain the value of the general concept of a Hilbert space we expound here a new method of solving the mixed problem for a homogeneous Laplace equation.

However, this method can also be applied to various other problems of mathematical physics. For simplicity we restrict ourselves to the case of two independent variables.

We consider a finite plane domain Ω; we shall assume that the boundary S of this domain is composed of a finite number of sufficiently smooth curves. We consider the set M of functions which are harmonic in Ω and continuous as far as the edge. We shall transform this set into a real Hilbert space by introducing scalar product according to the formula:

$$(u, v) = \int_S u(Q)v(Q)\mathrm{d}S; \qquad u, v \in M. \tag{1}$$

It is readily seen that in this case axioms A – D (§§ 3 and 42) are satisfied. It can be proved that the Hilbert space constructed is incomplete, we shall complete it; we denote the complete Hilbert space so obtained by H. If desired, it is possible to consider not functions harmonic in Ω but the limiting values of these functions on S as the elements of the space H; from this point of view space H is equivalent to

$L_2(S)$, the Hilbert space of functions which are quadratically summable on S.

We consider in the space H the operator

$$Au=\frac{\partial u}{\partial n}+\sigma(Q)u, \tag{2}$$

where $\sigma(P)$ is a function which is continuous and positive on S.

It is easily seen that operator (2) is positive-bounded-below. In fact, starting from Green's formula, it is easily verified that Au is a symmetric operator. Further

$$(Au, u)=\left(\frac{\partial u}{\partial n}, u\right)+(\sigma u, u).$$

By the well-known Green formula

$$\left(\frac{\partial u}{\partial n}, u\right)=\int_S u\frac{\partial u}{\partial n}\mathrm{d}S=\int_\Omega\left\{\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right\}\mathrm{d}\Omega+\int_\Omega u\nabla^2 u\,\mathrm{d}\Omega.$$

But u is a harmonic function, therefore $\nabla^2 u=0$ and

$$\left(\frac{\partial u}{\partial n}, u\right)=\int_\Omega\left\{\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)\right\}^2\mathrm{d}\Omega\geq 0. \tag{3}$$

Further

$$(\sigma u, u)=\int_S\sigma(Q)u^2(Q)\mathrm{d}S\geq\sigma_0\int_S u^2(Q)\mathrm{d}S=\sigma_0\|u\|^2,$$

where $\sigma_0=\min\ \sigma(Q)$; it follows from our assumptions that $\sigma_0>0$. Now, clearly,

$$(Au, u)\geq\sigma_0\|u\|^2,$$

i.e. operator (2) is positive-bounded-below. By virtue of the theorems of § 46, the problem of constructing a function which is harmonic in Ω and satisfies the condition

$$Au=\frac{\partial u}{\partial n}+\sigma(Q)u=f(Q), \tag{4}$$

on S is equivalent to the problem of the minimum of the functional

$$(Au, u)-2(u, f)=\left(\frac{\partial u}{\partial n}+\sigma u, u\right)-2(u, f).$$

This last problem has a solution which can be constructed by the Ritz method or as the limit of some other minimizing sequence.

We now explain the nature of the convergence of a minimizing sequence $u_n(P)$ to the exact solution $u(P)$. In all cases, by theorem 1 § 13, $\|u_n-u\|\to 0$. Further

$$u(P)-u_n(P)=\int_S [u(Q)-u_n(Q)]\frac{\partial G}{\partial n}\mathrm{d}S, \tag{5}$$

where G is Green's function (cf. 2, **198**) for the domain Ω and n is the external normal to S.

We consider a closed domain Ω' which lies entirely within Ω. If $P\in\Omega'$ and $Q\in S$, then the function $\dfrac{\partial G}{\partial n}$ remains bounded. Let $\dfrac{\partial G}{\partial n}<C$. (We can readily convince ourselves that $\dfrac{\partial G}{\partial n}$ is positive and we therefore omit the modulus symbol.) Now by Cauchy's inequality

$$|u(P)-u_n(P)|^2\le C^2\left[\int_S |u(Q)-u_n(Q)|\,\mathrm{d}S\right]^2$$

$$\le C^2|S|\int_S [u(Q)-u_n(Q)]^2\mathrm{d}S, \tag{6}$$

or

$$|u(P)-u_n(P)|\le C\gamma(|S|)\,\|u-u_n\|; \tag{7}$$

here $|S|$ is the length of the boundary S. It has been noted that $\|u-u_n\|\to 0$; now from inequality (7) it follows that *the approximate solutions tend to the exact one uniformly in any closed domain which lies entirely within* Ω.

Similarly it can be proved that the derivatives of any order of the approximate solutions tend uniformly to the corresponding derivative of the exact solution in any closed domain lying entirely within Ω.

CHAPTER VII

ESTIMATES OF THE ERROR OF AN APPROXIMATE SOLUTION

48. General remarks

As in the previous chapters we shall consider the equation

$$Au=f, \tag{1}$$

where A is a positive-bounded-below operator and f is a function with finite norm (or, in the more general case, an element of the Hilbert space in which operator A acts).

We have explained that equation (1) has a solution (which in the extreme case is a generalized one) and we have found a series of methods which permit this solution to be consstructed approximately. In the present chapter we shall consider the estimation of the error, which our approximate methods, especially the Ritz method, introduce.

In passing we shall be able to construct a new group of approximate methods. In this section we shall restrict ourselves to an indication of the fundamental principles upon which error estimation is based.

It is possible to estimate the error in an approximate solution simply but sometimes very crudely in the following way. Let u_n be an approximate solution of equation (1).

Knowing u_n, it is easy to evaluate the magnitude of $\|f-Au_n\|$. This characterizes the degree of precision with which the approximate solution satisfies the equation. The error in the approximate solution equals

$$u_0-u_n=GA(u_0-u_n)=G(f-Au_n).$$

By (11) § 45 we find an estimate for the norm of the error

$$\|u_0-u_n\|\leq\|G\|\,\|f-Au_n\|\leq\frac{1}{\gamma^2}\|f-Au_n\|. \tag{2}$$

Estimate (2) may be very crude since, in general, Au_n does not tend to f.

Another method of estimating the error in an approximate solution is as follows. We have

$$F(u_n)=|u_n-u_0|^2-|u_0|^2.$$

We use the notation

$$\min F(u)=-|u_0|^2=d.$$

Then

$$|u_n-u_0|^2=F(u_n)-d. \tag{3}$$

Usually we cannot determine the number d exactly. We suppose, however, that we have at our disposal a method which allows us to construct a number less than d and as close to d as required. Let δ be such a number. We replace d by δ in (3). This increases the right-hand side of (3) and we obtain the estimate sought

$$|u_n-u_0|<\gamma(F(u_n)-\delta). \tag{4}$$

From inequality (5) § 46 there also follows the estimate

$$\|u_n-u_0\|<\frac{1}{\gamma}\gamma(F(u_n)-\delta). \tag{5}$$

Thus the problem reduces to the construction of numbers δ less than d, and preferably as close as possible to d. Such a construction may be realized by using the method of orthogonal projections and the Trefftz method which are expounded in the following sections. We note, without going into the details, that the numbers δ can be constructed using the so called Friedrichs transformation which is expounded in sufficient detail in the monograph by R. Courant and D. Hilbert ([1], Ch. IV, § 9); further examples and some gene-

ralizations of the Friedrichs transformation can be found in the works of M.G. Slobodyansky [3–8].

The estimates (4) and (5) require the evaluation of the quantity $F(u_n)$. If u_n is constructed by the Ritz method, $F(u_n)$ can be evaluated in the following way.
We have:

$$F(u_n)=\sum_{k,\,m=1}^{n}(A\varphi_k,\varphi_m)a_k a_m-\sum_{k=1}^{n}a_k(\varphi_k,\ f)-\sum_{k=1}^{n}a_k(f,\ \varphi_k). \tag{6}$$

Further the coefficients a_k satisfy the equations

$$\sum_{k=1}^{n}(A\varphi_k,\ \varphi_m)a_k=(f,\ \varphi_m),\qquad m=1,\ 2,\ \ldots,\ n. \tag{7}$$

Multiplying this by a_m and summing over m, we obtain

$$\sum_{k,\,m=1}^{n}(A\varphi_k,\ \varphi_m)a_k a_m=\sum_{m=1}^{n}(f,\ \varphi_m)a_m=\sum_{k=1}^{n}(f,\ \varphi_k)a_k.$$

Substituting this in (6) we obtain

$$F(u_n)=-\sum_{k=1}^{n}a_k(\varphi_k,\ f)=-\sum_{k=1}^{n}a_k(f,\ \varphi_k). \tag{8}$$

Formula (8) is convenient because the numbers a_k and $(f,\ \varphi_k)$ which enter into it have been evaluated beforehand in the process of constructing and solving system (7). Formula (8) is derived on the assumption that all the functions considered are real.

In a complex Hilbert space, as is easily proved

$$F(u_n)=-\sum_{k=1}^{n}a_k\overline{(f,\ \varphi_k)}. \tag{9}$$

NOTE: As we know $d=F_{min}\ |u|=-\|u_0\|^2$. The problem thus consists of finding the negative numbers $\delta<-\|u_0\|^2$. This problem will be solved if we construct a positive functional $\Phi(u)$ the minimum (or at least the exact lower bound) of which equals $\|u_0\|^2$. In fact then we can put $\delta=-\Phi(u)$,

where u is any element from the field of definition of the functional $\Phi(u)$; obviously with a suitable choice of the element u we can ensure that δ is as close to $-\|u_0\|^2$ as we please.

§ 49. Subspaces and projections

We shall, as usual, consider the space $L_2(\Omega)$ of scalar or vectorial functions which are defined in a certain finite domain Ω and have a finite norm. We select some linear set of functions belonging to this class; we shall consider, without specially mentioning it, that all its limiting elements are added to this set (i.e. functions which are the limits in the mean of a sequence of functions belonging to the given linear set.) Such linear sets of functions are called *subspaces* of the basic space $L_2(\Omega)$.

Examples. 1. The set of functions which satisfy the equation

$$(u,\ 1)=\int_{\Omega} u\,d\Omega=0$$

is obviously linear and is the subspace of the functions whose mean values in the domain Ω equal zero.

2. The functions each of which is identically equal to some constant constitute a subspace.

3. Let Ω be equivalent to the segment $(0, 2\pi)$ of the x-axis.

The set of functions whose Fourier series contain only sines constitute a subspace; obviously the functions of this subspace are characterized by the fact that they are orthogonal to the functions $\cos nx$, $n=1, 2, \ldots$

4. We consider the class of vector functions which are defined in the finite domain Ω of three-dimensional space and possess finite norms in this domain.

The following are subspaces in this class:

α) The set of vectors whose divergence is zero;

β) the set of gradients of scalar functions;

γ) the set of gradients of scalar functions which vanish on the boundary of the domain.

Obviously, the examples cited in no way exhaust the manifold number of subspaces.

In the examples given the subspace constituted a part of

the fundamental class $L_2(\Omega)$: Sometimes the linear set leads to a subspace which coincides with $L_2(\Omega)$. Thus if we add to the set of all polynomials (which is obviously linear) all the limiting functions, then we obtain the class $L_2(\Omega)$. Another extreme case can arise where the subspace consists of only one function which is identically equal to zero. We shall not consider these two subspaces further.

Let there be given a separable (§ 44) Hilbert space H and one of its subspaces H_1. Obviously H_1 is also separable; by theorem 3 § 44 there exists in H_1 a complete finite or enumerable orthonormalized system. Let $\varphi_1(P)$, $\varphi_2(P)$, ..., $\varphi_n(P)$ be this system. We take an arbitrary function $u(P)$, which has a finite norm but which does not necessarily belong to the subspace H_1, and construct its orthogonal series in terms of the functions $\varphi_1(P), \varphi_2(P)$, ...:

$$\sum_{n=1}^{\infty} a_n \varphi_n(P), \qquad a_n = (u, \varphi_n). \tag{1}$$

The system $\varphi_1, \varphi_2, \ldots, \varphi_n, \ldots$ may be composed of a finite number of functions; in such a case the sum in (1) will contain only a finite number of terms.

By virtue of Bessel's inequality (§ 44), the series

$$\sum_{n=1}^{\infty} |a_n|^2$$

converges, but then series (1) also converges in the mean; its sum, which we denote by $u_1(P)$, belongs to the subspace H_1. In fact u_1 is the limit of the functions

$$\sum_{n=1}^{N} a_n \varphi_n(P),$$

which belong to the subspace H_1 as linear combinations of the functions $\varphi_n(P) \in H_1$. The function $u_1(P)$ is called *the orthogonal projection*, or simply *the projection* of the function u (P) in the subspace H_1.

We construct the difference $u_2(P) = u(P) - u_1(P)$. We shall

prove that the function $u_2(P)$ is orthogonal to any function from the subspace H_1. Firstly $(u_2, \varphi_k)=0$, $k=1, 2, \ldots$. In fact,

$$(u_2, \varphi_k)=(u, \varphi_k)-(u_1, \varphi_k)=a_k-\left(\sum_{k=1}^{\infty}a_n\varphi_n, \varphi_k\right)=a_k-a_k=0.$$

Further, if $v(P)$ is an arbitrary function from the subspace H_1 then it can be expanded as a Fourier series in the functions φ_n:

$$v(P)=\sum_{n=1}^{\infty}\alpha_n\varphi_n(P), \qquad \alpha_n=(v, \varphi_n).$$

Now

$$(u_2, v)=\sum_{n=1}^{\infty}\overline{\alpha_n}(u_2, \varphi_n)=0,$$

which was to be proved.

The fact that the function $u_2(P)$ is orthogonal to any function of the subspace H_1 is formulated briefly by saying that *the function $u_2(P)$ is orthogonal to the subspace H_1.*

From what has been proved it follows that any function $u(P)$ with finite norm can be represented as the sum of two terms, $u(P)=u_1(P)+u_2(P)$ of which the first is the projection of the given function $u(P)$ in the given subspace H_1 and the second term is orthogonal to this subspace.

Formula (1) defines the projection of the function $u(P)$ in the subspace H_1 in terms of a system of functions $\{\varphi_n(P)\}$ which is complete in H_1 and orthonormalized. Such a system is not unique and there arises the question of whether the projection does not change if the system $\{\varphi_n(P)\}$ is replaced by some other orthonormal system $\{\varphi_n(P)\}$ which is complete in H_1. We shall prove that this is not so and that the projection of a given function in a given subspace is defined uniquely. We suppose the converse to be true and let $u_1(P)$ and $\tilde{u}_1(P)$ be two projections of the function $u(P)$ onto the subspace H_1. The functions $u_2(P)=u(P)-u_1(P)$ and $\tilde{u}_2(P)=u(P)-\tilde{u}_1(P)$ are orthogonal to H_1. Then their difference is also orthogonal

to H_1. But $\tilde{u}_2(P)-u_2(P) = u_1(P)-u_1(P)$. The function $u_1(P)-\tilde{u}_1(P)$ belongs to H_1 as the difference of functions from this same subspace. Hence

$$0=(\tilde{u}_2-u_2,\ u_1-\tilde{u}_1)=(u_1-\tilde{u}_1,\ u_1-\tilde{u}_1)=\|u_1-\tilde{u}_1\|^2$$

and

$$u_1(P)\equiv\tilde{u}_1(P),$$

which was to be proved.

A projection possesses an important extremal property, namely, if $u(P)\in L_2(\Omega)$ and $u_1(P)$ is the projection of the function $u(P)$ onto the subspace H_1, then the norm of the difference $u(P)-v(P)$, where v is an arbitrary function from H_1, becomes a minimum for $v(P)\equiv u_1(P)$. In fact

$$\|u-v\|^2=\|u_2+(u_1-v)\|^2=\|u_2\|^2+\|u_1-v\|^2,$$

since $(u_1-v)\in H$ and thus $(u_2,\ u_1-v)=0$. It is clear now that $\|u-v\|^2$ becomes a minimum at $v=u_1$; this minimum equals $\|u_2\|^2$.

We consider all possible functions $u(P)\in L_2(\Omega)$ and their projections $u_1(P)$ in a given subspace H_1. The difference $u_2(P)=u(P)-u_1(P)$ as is easily seen, constitutes some new subspace H_2; any function from H_2 is orthogonal to any function from H_1, and it is said that the subspaces H_1 and H_2 are orthogonal. Further any function $u(P)\in L_2(\Omega)$ expands into a sum $u=u_1+u_2$, where $u_1\in H_1$ and $u_2\in H_2$. This fact is usually formulated by saying that $L_2(\Omega)$ is an *orthogonal sum of the subspaces H_1 and H_2*. It is also said that each of the subspaces H_1 and H_2 is *the orthogonal complement* of the other subspace.

All that we have said in the present section about $L_2(\Omega)$ applies without change to an arbitrary Hilbert space.

If a given Hilbert space is the orthogonal sum of the subspaces H_1 and H_2 then we write

$$H=H_1\oplus H_2.$$

This symbolic formula is also written in the form

$$H_1=H\ominus H_2;\quad H_2=H\ominus H_1.$$

The reader will easily notice the complete analogy between the theory of projections expounded in the present section and the theory of projections in ordinary three-dimensional space. Lines or planes passing through the coordinate origin are subspaces of three-dimensional space. If the given subspace is, for example, a straight line, then the plane normal to this straight line and passing through the coordinate origin is the subspace orthogonal to it; obviously three-dimensional space is the orthogonal sum of any straight line and of the plane normal to it, if both the line and the plane pass through the coordinate origin. In particular, for example, three-dimensional space is the orthogonal sum of the plane XOY and the axis OZ.

Any subspace, if it contains at least two linearly independent elements can, in its turn, be resolved into an orthogonal sum of two orthogonal subspaces; thus, the plane XOY is the orthogonal sum of the axes OX and OY. Thus, it is possible to arrive at the concept of the expansion of a given Hilbert space into an orthogonal sum of a (possibly infinite) number of orthogonal subspaces. Thus, three-dimensional space is the orthogonal sum of the axes OX, OY, OZ.

It is easily seen that the subspaces of examples 1 and 2 of this section are orthogonal and that their orthogonal sums give the space $L_2(\Omega)$. As we shall see below, subspaces α and γ of example 4 are also orthogonal; their orthogonal sum gives the whole of the space of vectors having a finite norm.

§ 50. The method of orthogonal projections in Dirichlet's problem

As has already been mentioned in the introduction, the method of orthogonal projections was formulated in 1909 by S. Zaremba in a paper of his [1] concerning the Dirichlet problem for the Laplace equation in three independent variables. While Zaremba could not make use of the theory of

operators, which at that time had not yet entirely been worked out, the basic characteristics of the method nonetheless emerged quite distinctly in the paper cited: Zaremba obtained the solution of the Dirichlet problem for the Laplace equation as the projection of a function which satisfies the posed boundary conditions in the subspace of harmonic functions. In a later paper [2] Zaremba applied the same method to the Dirichlet problem and to the Neumann problem for the Laplace equation in any number of independent variables. The term "method of orthogonal projections" itself first appeared in 1940 in a paper by H. Weyl [1] who also considered the Dirichlet problem; this method was applied to a wider class of boundary conditions by M. I. Vishik [1]–[3]. The works of J. Diaz [1] and K. Moren [1] are also concerned with the method of orthogonal projections. The further papers by M.I. Klyot-Dashinsky [1, 2] are also relevant.

We shall seek the function $u(P)$ which, within the finite domain Ω, satisfies the Poisson equation

$$-\nabla^2 = f(P), \tag{1}$$

and on the boundary S of the domain Ω, the boundary condition

$$u|_S = 0. \tag{2}$$

As usual, we shall consider that the function $f(P)$ has a finite norm. To be more explicit we shall suppose (although this is not essential) that Ω is a three-dimensional domain, so that

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}.$$

We use the notation $\nabla u = \mathbf{v}$. We note that our problem will be solved if we find the vector $\mathbf{v}$ since the setting up of a function of which it is the gradient is an extremely simple affair. Our problem can now be formulated thus: *it is required to find the vector* $\mathbf{v}(P)$ *which satisfies the equation*

$$-\nabla \cdot \mathbf{v} = f(P) \tag{3}$$

and is the gradient of some scalar function which vanishes on S.

We introduce the space of vector functions in which scalar product and norm are defined by formulae (11) and (12) § 3. For brevity we denote this space by $\mathfrak{H}$. We introduce in $\mathfrak{H}$ two subspaces which we denote by $\mathfrak{H}_1$, and $\mathfrak{H}_2$; as $\mathfrak{H}_1$, we take the subspace of vectors which are the gradients of scalar functions and are zero on S, as $\mathfrak{H}_2$ – the subspace of vectors which satisfy the equation

$$\nabla \cdot \mathbf{v} = 0. \tag{4}$$

These subspaces were mentioned in example 4 of the previous section. It should be borne in mind that subspaces $\mathfrak{H}_1$, and $\mathfrak{H}_2$ contain, besides the vectors described here, the limits of sequences of such vectors.

The following is founded upon the important formula

$$\mathfrak{H} = \mathfrak{H}_1 \oplus \mathfrak{H}_2. \tag{5}$$

Formula (5) contains two assertions which we are obliged to prove:

1) If the vector $\mathbf{v}_1$ is the gradient of some scalar which equals zero on S and the vector $\mathbf{v}_2$ has zero divergence, then these vectors are orthogonal in the sense that

$$(\mathbf{v}_1, \mathbf{v}_2) = \int_\Omega \mathbf{v}_1 \cdot \mathbf{v}_2 d\Omega = 0;$$

2) any vector with a finite norm can be represented as the sum of two vectors of which one is the gradient of a scalar vanishing on S and the other has zero divergence.

$$\text{Let } \mathbf{v}_1 = \nabla\varphi(P),\ \varphi(P)|_S = 0, \text{ and let } \nabla \cdot \mathbf{v}_2 = 0.$$

By the well-known formula of vector analysis

$$\mathbf{v}_1 \cdot \mathbf{v}_2 = \mathbf{v}_2 \cdot \nabla\varphi = \nabla \cdot (\varphi \mathbf{v}_2) - \varphi \nabla \cdot \mathbf{v}_2.$$

Integrating this equation over Ω and using Gauss's theorem in vector form (§ 2), we find

$$(\mathbf{v}_1, \mathbf{v}_2) = \int_S \varphi \mathbf{v}_2 \cdot d\mathbf{S} - \iiint_\Omega \varphi \nabla \cdot \mathbf{v}_2 d\Omega = 0,$$

since $\varphi=0$ on S and $\nabla\cdot\mathbf{v}_2=0$ in Ω. this proves assertion 1). We shall prove assertion 2) on the assumption that the given vector $\mathbf{v}$ has first derivatives which are continuous in $\bar{\Omega}=\Omega+S$. We denote by $\varphi(P)$ a function equal to zero on S and satisfying the equation $\nabla^2\varphi=\nabla\cdot\mathbf{v}$; the existence of such a function was proved in § 22. Now it suffices to put $\mathbf{v}_1=\nabla\varphi$, $\mathbf{v}_2=\mathbf{v}-\mathbf{v}_1$.

In the general case it is possible to construct a sequence of vectors $\mathbf{v}^{(n)}(P$ which have continuous first derivatives in $\bar{\Omega}$ and which tend to a given vector $\mathbf{v}$, so that

$$\|\mathbf{v}-\mathbf{v}^{(n)}\|\xrightarrow[n\to\infty]{}0. \tag{6}$$

Each of the vectors $\mathbf{v}^{(n)}$ can be broken down into the sum $\mathbf{v}^{(n)}=\mathbf{v}_1{}^{(n)}+\mathbf{v}_2{}^{(n)}$ where $\mathbf{v}_1{}^{(n)}\in\mathfrak{H}_1$ and $\mathbf{v}_2{}^{(n)}\in\mathfrak{H}_2$ Hence

$$\mathbf{v}^{(n)}-\mathbf{v}^{(m)}=(\mathbf{v}_1^{(n)}-\mathbf{v}_1^{(m)})+(\mathbf{v}_2^{(n)}-\mathbf{v}_2^{(m)}).$$

The terms on the right are orthogonal, therefore (§ 10)

$$\|\mathbf{v}^{(n)}-\mathbf{v}^{(m)}\|^2=\|\mathbf{v}_1^{(n)}-\mathbf{v}_1^{(m)}\|^2+\|\mathbf{v}_2^{(n)}-\mathbf{v}_2^{(m)}\|^2. \tag{7}$$

From (6) it follows that $\|\mathbf{V}^{(n)}-\mathbf{V}^{(m)}\|\xrightarrow[n,\,m\to\infty]{}0$, but then, by virtue of equation (7) we have that as $n\to\infty$ and $m\to\infty$ simultaneously

$$\|\mathbf{v}_1^{(n)}-\mathbf{v}_2^{(m)}\|\to 0;\quad \|\mathbf{v}_2^{(n)}-\mathbf{v}_2^{(m)}\|\to 0.$$

This shows that $\mathbf{v}_1{}^{(n)}$ and $\mathbf{v}_2{}^{(n)}$ tend in the mean to certain limits $\mathbf{v}_1$ and $\mathbf{v}_2$ respectively. Since $\mathbf{v}_1{}^{(n)}\in\mathfrak{H}_1$ and $\mathbf{v}_2{}^{(n)}\in\mathfrak{H}_2$ also $\mathbf{v}_1\in\mathfrak{H}_1$, $\mathbf{v}_2\in\mathfrak{H}_2$. Passing to the limit as $n\to\infty$ in the equation $\mathbf{v}^{(n)}=\mathbf{v}_1{}^{(n)}+\mathbf{v}_2{}^{(n)}$, we obtain

$$\mathbf{v}=\mathbf{v}_1+\mathbf{v}_2,$$

which was to be proved.

After these observations we pass on to the exposition of the method of orthogonal projections. We construct some vector $\mathbf{V}(P)$ which satisfies equation (3). In the given case this is quite easy: it is possible, for example, to take

$$V_x=V_y=0,\quad V_z=-\int f(x,\,y,\,z,)\mathrm{d}z.$$

We put $\mathbf{V}=\mathbf{v}+\mathbf{w}$ where $\mathbf{v}$ is the vector sought. Then

$$\nabla\cdot\mathbf{w}=\nabla\cdot\mathbf{V}-\nabla\cdot\mathbf{v}=0,$$

so that $\mathbf{w}\in\mathfrak{H}_2$. Now it is clear that *the vector sought is the projection of the vector* $\mathbf{V}$ *onto the subspace* $\mathfrak{H}_1$ and it is here that the essence of the method of orthogonal projections lies.

In § 49 a method was given for constructing projections which can be applied to determining the vector $\mathbf{v}$; it is easily seen that this method would lead to the orthogonal series of § 12. As we shall see below, *it proves expedient to project the vector* $\mathbf{V}$ *not into the subspace* $\mathfrak{H}_1$, *which is troublesome for us, but into the "complementary" subspace* $\mathfrak{H}_2$, which gives the vector $\mathbf{w}$; after this vector $\mathbf{v}$ is determined from the formula $\mathbf{v}=\mathbf{V}-\mathbf{w}$.

To construct the projection $\mathbf{w}$ we select a sequence of vectors $\psi(P)$ satisfying the equation $\nabla\cdot\psi_1=0$; if this sequence is orthonormal and complete in $\mathfrak{H}_2$, then by formula (1) § 49

$$\mathbf{w}=\sum_{n=1}^{\infty}a_n\psi_n,\quad a_n=(\mathbf{V},\psi_n),$$

and the solution of our problem is given by the formula

$$\mathbf{v}(P)=\mathbf{V}(P)-\sum_{n=1}^{\infty}(\mathbf{V},\psi_n)\psi_n(P). \tag{8}$$

We shall dwell a little upon an analysis of formula (8). We rewrite it in the form

$$\mathbf{V}(P)=\mathbf{v}(P)+\sum_{n=1}^{\infty}(\mathbf{V},\psi_n)\psi_n(P). \tag{9}$$

All the terms on the right in (9) are mutually orthogonal: the vectors $\psi_n(P)$ are orthogonal by definition, moreover $\mathbf{v}(P)$ and $\psi_n(P)$ are orthogonal, since they belong to the orthogonal subspaces $\mathfrak{H}_1$ and $\mathfrak{H}_2$. In this case by formulae (2) and (4) § 10

$$\|\mathbf{V}\|^2=\|\mathbf{v}\|^2+\sum_{n=1}^{\infty}(\mathbf{V},\psi_n)^2$$

and consequently

$$\|\mathbf{v}\|^2=\|\mathbf{V}\|^2-\sum_{n=1}^{\infty}(\mathbf{V},\psi_n)^2. \tag{10}$$

If in the series on the right-hand side of (10) only the first m terms are retained,

$$\|\mathbf{v}\|^2\leq\|\mathbf{V}\|^2-\sum_{n=1}^{m}(\mathbf{V},\psi_n)^2. \tag{11}$$

This corresponds to replacing the exact solution (8) by an approximate solution according to the formula

$$\mathbf{v}(P)\approx\mathbf{v}^{(m)}(P)=\mathbf{V}(P)-\sum_{n=1}^{m}(\mathbf{V},\psi_n)\psi_n(P).$$

By the definition of the norm of a vector function [formula (12) § 3]

$$\|\mathbf{v}\|^2=\int_{\Omega}(v_x^2+v_y^2+v_z^2)\,d\Omega.$$

Denoting by $u_0(P)$ a function which satisfies equations (1) and (2), we have $\mathbf{v}=\nabla u_0$ and

$$\|\mathbf{v}\|^2=\int_{\Omega}\left[\left(\frac{\partial u_0}{\partial x}\right)^2+\left(\frac{\partial u_0}{\partial y}\right)^2+\left(\frac{\partial u_0}{\partial z}\right)^2\right]d\Omega,$$

or, by formula (4) § 22

$$\|\mathbf{v}\|^2=(-\nabla^2u_0,\,u_0)=\|u_0\|^2. \tag{12}$$

On the other hand, by formula (4) § 11,

$$\|u_0\|^2=-\min F(u), \tag{13}$$

where $F(u)=(-\nabla^2u,\,u)-2(u,\,f)$ is the functional used in the energy method. Now from formulae (11)–(13) it follows that

$$\min F(u)\geq-\left\{\|\mathbf{V}\|^2-\sum_{n=1}^{m}(\mathbf{V},\,\psi_n)^2\right\}. \tag{14}$$

Now returning to § 48, we see that *the method of orthogonal projections allows us to estimate the error in an approximate solution constructed by the Ritz method.* In fact, inequality (14) shows that it is possible to put

$$\delta=-\|\mathbf{V}\|^2+\sum_{n=1}^{m}(\mathbf{V},\,\psi_n)^2; \tag{15}$$

in formulae (4) and (5) § 48, in this case, as is clear from formulae (10), (12) and (13),

$$\lim_{m\to\infty}\delta=d=\min F(u).$$

We shall again return to the question of the estimation of error in the next section.

§ 51. General formulation of the method of orthogonal projections

After the Dirichlet problem considered in § 50 it will be simple to give a more general formulation of the method of orthogonal projections, a formulation which allows us to apply this method to a number of other boundary value problems.

Suppose we require to solve equation

$$Au=f, \tag{1}$$

where the given element f and the element sought u both belong to some Hilbert space H, and the operator A is positive-bounded-below in this space. We further suppose that the given problem has led to the seeking of the element v of some (in general) new Hilbert space $\mathfrak{H}$, and let the element v satisfy the equation

$$Bv=f, \tag{2}$$

in which B is some linear operator.

We consider the set of solutions of the homogeneous equation

$$Bv=0. \tag{3}$$

This set is linear; by the addition to it of its limiting elements which we shall take as generalized solutions of equation (3), we obtain a subspace of the space $\mathfrak{H}$; we shall denote this subspace by $\mathfrak{H}_2$. We make one further assumption which is important for the method of orthogonal projections; we assume that *the element v sought belongs to a subspace $\mathfrak{H}_1$ which is orthogonal to $\mathfrak{H}_2$.*

Equation (2) is solved by the method of orthogonal pro jections thus: we construct some element V which satisfies the equation $BV=f$ and we put $V-v=w$. Then $Bw=BV-Bv=0$; this implies that w is an element of subspace $\mathfrak{H}_2$; the equation $V=v+w$ shows that v is the orthogonal projection of the element V on the subspace $\mathfrak{H}_2$.

The construction of the element v can be carried out in two ways. We can construct a system of elements φ_n, $n=1, 2, \ldots$, which is orthonormal and complete in $\mathfrak{H}_1$; then

$$v=\sum_{n=1}^{\infty}(V, \varphi_n)\varphi_n. \tag{4}$$

Usually formula (4) is identical to formula (2) § 12 which gives the solution of equation (1). Thus, this variant of the method of orthogonal projections essentially coincides with the Ritz method. Alternatively we can construct the element v sought by choosing an orthonormal system ψ_n, $n=1, 2, \ldots$, which is *complete in* $\mathfrak{H}_2$. Then

$$w=\sum_{n=1}^{\infty}(V, \psi_n)\psi_n$$

and

$$v=V-\sum_{n=1}^{\infty}(V, \psi_n)\psi_n. \tag{5}$$

In the case of the Dirichlet problem considered in the previous section H is the subspace $L_2(\Omega)$ of functions which are quadratically summable in the given domain Ω, and $\mathfrak{H}$ is the space of vector functions which also are quadratically summable in Ω. Further $B\mathbf{v}=-\nabla.\mathbf{v}$; $\mathfrak{H}_2$ is the subspace of vectors with zero divergence and the subspace $\mathfrak{H}_1$, orthogonal to $\mathfrak{H}_2$, consists as was explained in § 50, of the gradients of scalar functions which vanish on the boundary S of the domain.

If in series (5) we retain only a finite number m of terms, then the method of orthogonal projections leads to the approximate solutions

$$v_m=V-\sum_{n=1}^{m}(V, \psi_n)\psi_n. \tag{6}$$

We now make a useful practical observation. In order to construct the approximate solution (6) it suffices to have m normalized elements $\psi_1, \psi_2, \ldots, \psi_m$ which are mutually orthogonal and belong to the subspace $\mathfrak{H}_2$. We now suppose that we have constructed m not necessarily orthonormal but linearly independent elements $\omega_1, \omega_2, \ldots, \omega_m$ of this subspace. The process of orthogonalization (see § 10) enables us to transform the elements $\omega_1, \omega_2, \ldots, \omega_n$ into the orthonormalized

elements $\psi_1, \psi_2, \ldots, \psi_m$ and then to construct an approximate solution of the problem by formula (6). It so happens that the process of orthogonalization can be avoided.

We consider the expression

$$V - \sum_{n=1}^{m} \alpha_n \psi_n \tag{7}$$

and we pose the problem of selecting coefficients α_n such that the norm of the expression (7) is a minimum. Since the elements ψ_n are orthogonal and normalized, we find by repeating the reasoning of § 10, that $\alpha_n = (V, \psi_n)$. Thus the element v_m [formula (6)] solves the problem of minimizing the quantity

$$\|V - \sum_{n=1}^{m} \alpha_n \psi_n\|.$$

As follows from the process of orthogonalization the elements $\psi_1, \psi_2, \ldots, \psi_n$ can be linearly expressed in terms of $\omega_1, \omega_2, \ldots, \omega_m$ and vice versa. Hence it follows that v_m can be determined as the element minimizing the quantity.

$$\|V - \sum_{n=1}^{m} b_n \omega_n\|, \tag{8}$$

where $b_1, b_2, \ldots, b_m$ are constants; the problem of the minimum of the quantity (8) is easily solved.

Firstly instead of the minimum of quantity (8) it is possible to determine the minimum of the quantity:

$$\|V - \sum_{n=1}^{m} b_n \omega_n\|^2 - \|V\|^2 = (V - \sum_{n=1}^{m} b_n \omega_n, V - \sum_{n=1}^{m} b_n \omega_n) - (V, V). \tag{9}$$

Quantity (9) is a function of the independent variables $b_1, b_2, \ldots, b_m$ In order to find the minimum of this function we equate its partial derivatives with respect to these variables to zero. Repeating the reasoning of § 14, we obtain the system

$$\sum_{n=1}^{m} b_n (\omega_n, \omega_k) = (V, \omega_k), \quad k = 1, 2, \ldots, n. \tag{10}$$

Finding the b_n from this system, we obtain v_m in the form

$$v_m = V - \sum_{n=1}^{m} b_n \omega_n. \tag{11}$$

The construction and solution of system (10) is generally less troublesome than the process of orthogonalization.

NOTE: System (10) still retains its form in a complex Hilbert space.

In the last paragraph we saw that the method of orthogonal projections allows us to estimate an approximate solution constructed by the Ritz method. In the general case such an estimate is based upon the additional assumption which is satisfied for an extremely large number of cases:

The element u_0 satisfying equation (1), *and the element finding which in the method of orthogonal projections corresponds to finding the element u, are connected by the relation*

$$\|v\| = \|u_0\|; \tag{12}$$

the norm of the element v being evaluated in the space $\mathfrak{H}$.

Supposing that the relation (12) is satisfied, the rest of the argument proceeds as in § 50. We have

$$\|v\|^2 = \|V\|^2 - \sum_{n=1}^{\infty} (V, \psi_n)^2 \quad \text{and} \quad \|v_m\|^2 = \|V\|^2 - \sum_{n=1}^{m} (V, \psi_n)^2.$$

Hence it follows that

$$\|v\|^2 \leq \|v_m\|^2 \tag{13}$$

and

$$\lim_{m \to \infty} \|v_m\|^2 = \|v\|^2. \tag{14}$$

Putting

$$F(u) = (Au, u) - 2(u, f),$$

we have by formula (4), § 11 and the relations (12) and (13) that

$$\min F(u) = F(u_0) = -\|v\|^2 \geq -\|v_m\|^2.$$

In formulae (4) and (5), § 48 it is now possible to put

$$\delta = -\|v_m\|^2 = -\|V\|^2 + \sum_{n=1}^{m} (V, \psi_m)^2, \tag{15}$$

so that

$$\lim_{m\to\infty} \delta = d = \min F(u).$$

We further recall that by formula (12) § 14 $F(u_n) = -\|u_n\|^2$ where u_n is Ritz's approximate solution of equation (1). Formula (4) § 48 now gives an estimate of the error in the Ritz method:

$$\|u-u_n\| \leq \gamma(\|v_m\|^2 - \|u_n\|^2). \tag{16}$$

Similarly, by virtue of formula (5) § 48

$$\|u-u_n\| \leq \frac{1}{\gamma}\gamma(\|v_m\|^2 - \|u_n\|^2). \tag{17}$$

We note that by formula (8) § 48

$$\|u_n\|^2 = \sum_{k=1}^{n} a_k(f, \varphi_k).$$

Similarly we find

$$\|v_m\|^2 = \|V\|^2 - \sum_{k=1}^{m} b_k(V, \omega_k). \tag{18}$$

It is also easy to give an estimate to the error of the approximate solution (6) constructed by the method of orthogonal projections. We put

$$w = \sum_{n=1}^{\infty} (V, \psi_n)\psi_n, \qquad w_m = \sum_{n=1}^{m} (V, \psi_n)\psi_n.$$

Firstly we have

$$v - v_m = w_m - w \text{ and } \|v-v_m\| = \|w-w_m\|.$$

Further, $V = v + w$, hence

$$v_m = V - w_m = v + (w - w_m).$$

The terms on the right are orthogonal, since $v \in \mathfrak{H}_1$ and $(w-w_n) \in \mathfrak{H}_2$.

Hence

$$\|v_m\|^2 = \|v\|^2 + \|w-w_m\|^2$$

and

$$\|v-v_m\|^2 = \|w-w_m\|^2 = \gamma(\|v_m\|^2 - \|v\|^2).$$

As before, let u_n denote the Ritz approximate solution of equation (1), then

$$\|u_n\|^2 \leq \|u_0\|^2 = \|v\|^2.$$

This last equation gives the estimate sought

$$\|v - v_m\| \leq \gamma(\|v_m\|^2 - \|u_n\|^2). \tag{19}$$

It is relevant to observe that in the estimates (16) and (19) the right-hand sides are identical.

In Ch. VIII numerical examples of the estimation of error using the method of orthogonal projections will be given.

In concluding this section we make the following observation. The element v, being a projection of the element V has the least norm; since V is an arbitrary element satisfying equation (2) the element v is the solution of the following variational problem: *to find the solution of the equation $Bv=f$ which has the lowest norm.*

§ 52. Some additional considerations

In the previous section a discussion was given of the method for othogonal projections which was based upon the following assumptions:
(a) the solution of the equation $Au=f$ reduces to the seeking of some element v which belongs to a new Hilbert space $\mathfrak{H}$ and satisfies the equation $Bv=f$;
(b) the element v belongs to the subspace which is orthogonal to the subspace of the solutions of the homogeneous equation $Bv=0$;
(c) $\|v\| = \|u\|$.

In the present section which assumes a knowledge of the fundamentals of functional analysis we shall indicate rather more general conditions the fulfilment of which necessitates the fulfilment of the assumptions enumerated above. This condition consists of the following: *a positive definite operator A can be represented in the form of a product of two conjugate operators*

$$A = T^*T, \tag{1}$$

the operator T transforming from a given Hilbert space H into some Hilbert space $\mathfrak{H}$, and being in particular, defined, for the entire space H_A.

We note that in this case the conjugate operator T^* transforms from the space $\mathfrak{H}$ into the space H. The equation $Au=f$ now assumes the form

$$T^*Tu = f. \tag{2}$$

We put $Tu=v$. Then

$$T^*v=f. \tag{3}$$

This justifies the assumption (a) if we put $B=T^*$. The case where the homogeneous equation

$$T^*u=0 \tag{4}$$

has solutions different from zero is of practical interest. The set of these olutions constitutes some subspace $\mathfrak{H}_2$ of the space $\mathfrak{H}$. We shall prove that the element, $v=Tu$, sought is orthogonal to $\mathfrak{H}_2$. In fact if $w\in\mathfrak{H}_2$, then $T^*w=0$ and

$$(v,\ w)_{\mathfrak{H}}=(Tu,\ w)_{\mathfrak{H}}=(u,\ T^*w)_H=0 \tag{5}$$

(the suffix indicates the space in which the scalar product is evaluated). Equation (5) proves the validity of the assumption (b).

We now turn to the assumption (c). Let $u_1\in D_A$, $u_2\in D_A$. Then, putting $Tu_1=v_1$ and $Tu_2=v_2$ we have

$$[u_1,\ u_2]=(Au_1,\ u_2)_H=(T^*Tu_1,\ u_2)_H=(Tu_1,\ Tu_2)_{\mathfrak{H}}=(v_1,\ v_2)_{\mathfrak{H}}.$$

In particular, if $u_1=u_2=u$, $Tu=v$ then

$$\|u\|=\|v\|_{\mathfrak{H}}. \tag{6}$$

Equation (6) may be established for all $u\in H_A$ by means of a limiting process, and thus the assumption (c) holds. We also note the equality of the scalar products

$$[u_1,\ u_2]=(Tu_1,\ Tu_2)_{\mathfrak{H}}. \tag{7}$$

As soon as the assumptions (a), (b), (c) have been established, further construction is extremely simple. Let $\mathfrak{H}_1$ be the subspace orthogonal to $\mathfrak{H}_2$ and let V be some solution of equation (3). Then the element v sought is the projection of the element V in the subspace $\mathfrak{H}_1$. Having constructed the element v, it remains to find u from the equation $Tu=v$. From formula (6) [assumption (c)] it follows that the operator T^{-1} is defined for the entire subspace $\mathfrak{H}_1$, and that its norm, as an operator transforming from $\mathfrak{H}_1$ to $\mathfrak{H}_A$, equals unity. In practical problems the operator T^{-1} is usually very simple and the equation $Tu=v$ is solved without difficulty (we met an example of this kind in § 50).

We now prove that the direct projection in the space $\mathfrak{H}_1$, leads to formulae 2 § 12 and consequently is equivalent to the application of the Ritz method.

Let $\{\eta_j\}$ be a system, complete and orthonormal in $\mathfrak{H}_1$, so that

$$(\eta_j,\ \eta_k)_{\mathfrak{H}}=\begin{cases}0, & j\neq k,\\ 1, & j=k;\end{cases} \tag{8}$$

then

$$v=\sum_{j=1}^{\infty}(V,\ \eta_j)\eta_j. \tag{9}$$

We put $T^{-1}\eta_j=\varphi_j$. It is easily seen that the sequence $\{\varphi_j\}$ is orthonormal and complete in H_A. In fact

$$(\eta_j,\ \eta_k)_{\mathfrak{H}}=(T\varphi_j,\ T\varphi_k)_{\mathfrak{H}},$$

and by formulae (7) and (8)

$$[\varphi_j, \varphi_k]=\begin{cases} 0, & j\neq k, \\ 1, & j=k. \end{cases}$$

We suppose that the sequence $\{\varphi_j\}$ is not complete in H_A. Then there exists an element $\varphi\in H_A$ such that $[\varphi, \varphi_j]=0$, $j=1, 2, \ldots$.
By formula (7), $(T\varphi, \eta_j)_{\mathfrak{H}}=0$, $j=1, 2, \ldots$. Since the system $\{\eta_j\}$ is complete in $\mathfrak{H}_1$, $T\varphi\in\mathfrak{H}_2$. But then $T^*(T\varphi)=0$ or $A\varphi=0$ and since A is positive definite, $\varphi=0$, contrary to assumption.

Acting on both sides of equation (9) with operator T which is bounded in $\mathfrak{H}_1$, we have

$$u=\sum_{j=1}^{\infty}(V, \eta_j)_{\mathfrak{H}}\varphi_j, \tag{10}$$

Further, $V=v+w$, $w\in\mathfrak{H}_2$. Hence

$$(w, \eta_j)_{\mathfrak{H}}=0$$

and

$$(V, \eta_j)_{\mathfrak{H}}=(v, \eta_j)_{\mathfrak{H}}=(Tu, T\varphi_j)_{\mathfrak{H}}=[u, \varphi_j]=(Au, \varphi_j)_H=(f, \varphi_j)_H.$$

Formula (10) assumes the form

$$u=\sum_{j=1}^{\infty}(f, \varphi_j)_H\varphi_j,$$

which is identical with formula (2) § 12.

§ 53. The Neumann problem

The Neumann problem for the Poisson equation consists of finding a function $u(P)$ which satisfies the equation

$$-\nabla^2 u=f(P) \tag{1}$$

within the domain Ω, and the boundary condition

$$\left.\frac{\partial u}{\partial n}\right|_S=0 \tag{2}$$

on its boundary S. A solution exists if the function $f(P)$, the norm of which is assumed to be finite, is orthogonal to unity, so that

$$\int_{\Omega} f(P)\,\mathrm{d}\Omega=0. \tag{3}$$

The solution is unique if we subject it to the requirement of orthogonality to unity:

$$\int_{\Omega} u(P)\mathrm{d}\Omega = 0. \tag{4}$$

With a view to the application of the method of orthogonal projections, we take as the space H the set of functions which have a finite norm and are orthogonal to unity and as the space $\mathfrak{H}$, the set of vectors with scalar product and norm defined by formulae (11) and (12) § 3. We put $\nabla u = \mathbf{v}$; equations (1) and (2) then give

$$-\nabla\cdot\mathbf{v} = f(P), \quad v_n|_S = 0. \tag{5}$$

Thus in our problem the operator B is the operator $-\nabla$ when applied to vectors whose normal components vanish on S. The subspace $\mathfrak{H}_2$ consists of the vectors which satisfy the equations

$$\nabla\cdot\mathbf{w} = 0, \quad w_n|_S = 0. \tag{6}$$

By repeating the reasoning of § 50 we satisfy ourselves that the orthogonal complement of $\mathfrak{H}_2$ is the subspace, $\mathfrak{H}_1$, of the gradients of *all possible* scalar functions.

If we succeed in finding a vector $\mathbf{V}$ such that $-\nabla\cdot\mathbf{V} = f$ and $V_n|_S = 0$ then $\mathbf{v}$ is found as the projection of the vector $\mathbf{V}$ in H_1. For its construction we must select a sequence (complete in $\mathfrak{H}_2$) of vectors $\mathbf{w}^{(n)}(P)$ which we shall consider to be orthonormal in order to simplify the argument; then

$$\mathbf{v} = \nabla u = \mathbf{V} - \sum_{n=1}^{\infty} (\mathbf{V}, \mathbf{w}^{(n)})\mathbf{w}^{(n)}(P).$$

The coordinate vectors $\mathbf{w}^{(n)}$ must satisfy equations (6); the actual construction of such a system meets with certain practical difficulties.

NOTE: The method of orthogonal projections can be extended without any difficulty to elliptic equations of the more general form

$$-\sum_{j,k=1}^{m} \frac{\partial}{\partial x_j}\left(A_{jk}\frac{\partial u}{\partial x_k}\right) = f(P). \tag{7}$$

In this case as $\mathfrak{H}$ we can take the space of vectors $\mathrm{v}(v_1, v_2, \ldots, v_m)$ with the scalar product

$$(\mathrm{v}, \mathrm{w}) = \int_{\Omega} \sum_{j,k=1}^{m} C_{jk} v_j w_k \mathrm{d}\Omega, \tag{8}$$

where the matrix $\|C_{jk}\|_{j,k=1}^{j,k=m}$ is the inverse of the matrix $\|A_{jk}\|_{j,k=1}^{j,k=m}$. For example consider the condition $u|_s = 0$.
As the subspace H_1 we take the set of vectors $\mathrm{v}(v_1, v_2, \ldots, v_m)$ where

$$v_j = \sum_{k=1}^{m} A_{jk} \frac{\partial u}{\partial x_k} \tag{9}$$

and $u(P)$ is a function equal to zero on S. It is readily proved that the orthogonal subspace $\mathfrak{H}_2 = \mathfrak{H} \ominus \mathfrak{H}_1$, consists of vectors satisfying the equation

$$\nabla \cdot \mathrm{v} = \sum_{j=1}^{m} \frac{\partial v_j}{\partial x_j} = 0. \tag{10}$$

If the boundary condition

$$\left[\sum_{j,k=1}^{m} A_{jk} \frac{\partial u}{\partial x_k} \cos(n, x_j) \right]_S = 0, \tag{11}$$

is posed, then $\mathfrak{H}_1$ is constructed as the totality of the vectors (9), but this time no boundary condition is imposed upon the function $u(P)$. In this case the orthogonal complement $\mathfrak{H}_2 = \mathfrak{H} \ominus \mathfrak{H}_1$, consists of vectors which satisfy equation (10) and moreover, the boundary condition

$$v_n|_S = 0. \tag{12}$$

The construction of the space $\mathfrak{H}$ and its expansion into orthogonal subspaces becomes more involved if a term of the type $C(P)u(P)$ is added to the left-hand side of equation (7); an example of such a type has been worked out in a paper by M.I. Vishik [1]. In other papers by the same author [2]–[3] various examples of equations of higher order are considered.

§ 54. Castigliano's principle and two-sided estimates in elasticity theory

The well-known principle in elasticity theory due to Castigliano can be obtained by starting from the method of orthogonal projections. In order to simplify notation we shall restrict ourselves to the case of a homogeneous isotropic elastic medium, although the general case of a non-homogeneous and anisotropic medium introduces no essential complications. We shall also assume that the boundary conditions are homogeneous.

We consider some body Ω, bounded by the surface S. As usual we shall consider that the body is finite and that its bounding surface is piecewise smooth. We shall consider all possible stress tensors, defined in $\overline{\Omega}=\Omega+S$. We shall denote such a tensor by T, and its components by $\sigma_x, \sigma_{xy}, \ldots, \tau_{yz}$:

$$T=\begin{pmatrix} \sigma_x, & \tau_{xy}, & \tau_{xz} \\ \tau^x{}_y, & \sigma_y, & \tau_{yz} \\ \tau_{xz}, & \tau_{yz}, & \sigma_z \end{pmatrix}. \tag{1}$$

With each tensor T we associate a tensor $E(T)$ according to the formulae

$$E(T)=\begin{pmatrix} \varepsilon_x, & \gamma_{xy}, & \gamma_{xz} \\ \gamma_{xy}, & \varepsilon_y, & \gamma_{yz} \\ \gamma_{xz}, & \gamma_{yz}, & \varepsilon_z \end{pmatrix}; \tag{2}$$

$$\begin{aligned} \varepsilon_x&=\frac{1}{E}[\sigma_x-\sigma(\sigma_y+\sigma_z)], \quad \gamma_{xy}=\frac{2(1+\sigma)}{E}\tau_{xy}, \\ \varepsilon_y&=\frac{1}{E}[\sigma_y-\sigma(\sigma_x+\sigma_z)], \quad \gamma_{xz}=\frac{2(1+\sigma)}{E}\tau_{xz}, \\ \varepsilon_z&=\frac{1}{E}[\sigma_z-\sigma(\sigma_x+\sigma_y)], \quad \gamma_{yz}=\frac{2(1+\sigma)}{E}\tau_{yz}. \end{aligned} \tag{3}$$

Here E is Young's modulus and σ Poisson's ratio for the material of the body Ω. If the body has not been strained beyond the elastic limit then $\varepsilon_x, \varepsilon_y, \ldots \gamma_{xy}$ are the deformation components: they are related to the displacement components

by formulae (1) § 26. Equations (3) are equivalent to the Lamé equations (§ 26).

We transform the set of tensors T into a real Hilbert space by defining scalar product and norm in it by the formulae:

$$(T',T'')=\int_{\Omega}(\sigma_x'\varepsilon_x''+\sigma_y'\varepsilon_y''+\sigma_z'\varepsilon_z''+\tau_{xy}'\gamma_{xy}''+\tau_{xz}'\gamma_{xz}''+\tau_{yz}'\gamma_{yz}'')\,\mathrm{d}\Omega; \quad (4)$$

$$\|T\|^2=\int_{\Omega}(\sigma_x\varepsilon_x+\sigma_y\varepsilon_y+\sigma_z\varepsilon_z+\tau_{xy}\gamma_{xy}+\tau_{xz}\gamma_{xz}+\tau_{yz}\gamma_{yz})\,\mathrm{d}\Omega. \quad (5)$$

More precisely the Hilbert space is formed by the totality of the tensors for which integral (5) is finite.

Using formulae (3) it is readily proved that the scalar product as defined by formula (4) satisfies axioms A–D § 3.

In particular, it is readily proved that

$$\sigma_x'\varepsilon_x''+\ldots+\tau_{yz}'\gamma_{yz}''=\sigma_x''\varepsilon_x'+\ldots\tau_{yz}''\gamma_{yz}';$$

and so the symmetry of the scalar product follows. We denote the Hilbert space thus constructed by $\mathfrak{H}$.

We note that if T is the tensor of the *elastic* stresses then $\frac{1}{2}\|T\|^2$ is the potential energy of deformation of the body Ω. Extending the definition, we shall call the quantity $\frac{1}{2}\|T\|^2$ the potential energy of deformation for any stress tensor T.

We shall consider the mixed problem of elasticity theory (problem (c) § 26); from it, as particular cases, we obtain problem (a) concerning a body with a fixed boundary and problem (b) concerning a body with a free boundary. We recall that the mixed problem of elasticity theory consists in determining the vector of the elastic displacement $\mathbf{u}(u_x, u_y, u_z)$ and of the deformations and stresses connected with it; the quantities mentioned are interrelated by equations (3) (or by the Lamé equations) and by equations (1) § 26. Further the stress components satisfy the equilibrium equation (see equation (5) § 26)

$$\left.\begin{aligned} -\frac{\partial\sigma_x}{\partial x}+\frac{\partial\tau_{xy}}{\partial y}+\frac{\partial\tau_{xz}}{\partial z} &= X,\\ -\frac{\partial\tau_{xy}}{\partial x}+\frac{\partial\sigma_y}{\partial y}+\frac{\partial\tau_{yz}}{\partial z} &= Y,\\ -\frac{\partial\tau_{xz}}{\partial x}+\frac{\partial\tau_{yz}}{\partial y}+\frac{\partial\sigma_z}{\partial z} &= Z, \end{aligned}\right\} \tag{6}$$

where X, Y, Z are the components of $\mathbf{K}$ the force per unit volume. Finally, the boundary conditions are formulated in the following fashion: the surface S is divided into two parts S_1 and S_2, where

$$\mathbf{u}|_{S_1}=0 \tag{7}$$

and

$$\mathbf{t}^{(n)}|_{S_2}=0; \tag{8}$$

$\mathbf{t}^{(n)}$ is the stress vector acting on the surface element with normal n.

We consider the set of stress tensors which satisfy boundary conditions (8) and the homogeneous equilibrium equations

$$\left.\begin{aligned} \frac{\partial\sigma_x}{\partial x}+\frac{\partial\tau_{xy}}{\partial y}+\frac{\partial\tau^x{}_z}{\partial z}&=0,\\ \frac{\partial\tau_{xy}}{\partial x}+\frac{\partial\sigma_y}{\partial y}+\frac{\partial\tau_{yz}}{\partial z}&=0,\\ \frac{\partial\tau_{xz}}{\partial x}+\frac{\partial\tau_{yz}}{\partial y}+\frac{\partial\sigma_z}{\partial z}&=0. \end{aligned}\right\} \tag{9}$$

This set, to which must be added its limiting elements (in the sense of convergence in space $\mathfrak{H}$), constitutes in $\mathfrak{H}$ a subspace which we denote by $\mathfrak{H}_2$. We note that in the case of problem (a), where the entire surface of the body is fixed, condition (8) is absent and the tensors entering into the subspace $\mathfrak{H}_2$ need satisfy no boundary conditions.

As the space $\mathfrak{H}_1$, we take the set (to which must be added its limiting elements, in the sense of convergence in space $\mathfrak{H}$) of stress tensors possessing the property that the quantities ε_x, ε_y, . . . , γ_{yz}, corresponding to them are deformations asso-

ciated with a certain vector $\mathbf{u}(P)$ by formulae (1) § 26. We further require that the vector $\mathbf{u}(P)$ satisfy boundary condition (7). In the case of problem (b), where the boundary of the body is stress-free, boundary condition (7) is absent and the vectors $\mathbf{u}(P)$ with which the tensors from the subspace $\mathfrak{H}_1$ are associated need satisfy no boundary conditions.

The whole basis of the construction of a solution is the equation

$$\mathfrak{H} = \mathfrak{H}_1 \oplus \mathfrak{H}_2. \tag{10}$$

The proof of equation (10) reduces to establishing two facts:

(1) If $T' \in \mathfrak{H}_1$, and $T'' \in \mathfrak{H}_2$ then $(T', T'') = 0$. We have

$$(T', T'') = \int_{\Omega} (\varepsilon'_x \sigma''_x + \varepsilon'_y \sigma''_y + \varepsilon'_z \sigma''_z + \gamma'_{xy} \tau''_{xy} + \gamma'_{xz} \tau''_{xz} + \gamma'_{yz} \tau''_{yz}) \mathrm{d}\Omega.$$

The quantities $\varepsilon'_x, \ldots, \gamma'_{yz}$ are connected by formulae (1) § 26 with a certain vector $\mathbf{u} = (u_x, u_y, u_z)$. Using these formulae and then integrating by parts, we obtain

$$(T', T'') = \int_{\Omega} \left\{ u'_x \left(\frac{\partial \sigma''_x}{\partial x} + \frac{\partial \tau''_{xy}}{\partial y} + \frac{\partial \tau''_{xz}}{\partial z} \right) + u'_y \left(\frac{\partial \tau''_{xy}}{\partial x} + \frac{\partial \sigma''_y}{\partial y} + \frac{\partial \tau''_{yz}}{\partial z} \right) \right.$$
$$\left. + u'_z \left(\frac{\partial \tau''_{xz}}{\partial x} + \frac{\partial \tau''_{yz}}{\partial y} + \frac{\partial \sigma''_z}{\partial z} \right) \right\} \mathrm{d}\Omega + \int_S \mathbf{u}' \cdot \mathbf{t}''^{(n)} \mathrm{d}S.$$

Since tensor T'' belongs to subspace $\mathfrak{H}_2$ its components satisfy the homogeneous equilibrium equations (9) and the volume integral vanishes. The surface integral also vanishes, since $\mathbf{u}'|_{S_1} = 0$, $\mathbf{t}''^{(n)}|_{S_2} = 0$. Finally $(T', T'') = 0$ and the subspaces $\mathfrak{H}_1$ and $\mathfrak{H}_2$ are orthogonal.

2) If T is an arbitrary tensor from $\mathfrak{H}$, then it can be represented in the form $T = T' + T''$, where $T' \in \mathfrak{H}_1$ and $T'' \in \mathfrak{H}_2$. We carry out the proof on the assumption that the components of the tensor T are continuous and continuously differentiable in $\bar{\Omega} = \Omega + S$; the extension to the general case is carried out as in § 50. We denote by $\mathbf{u}'$ the elastic displacement vector satisfying, together with the corresponding stress tensor T', equations (1) § 26 and also the equations (3) and (6) and the

boundary conditions (7) and (8). It now suffices to put $T''=$ $=T-T'$.

We now turn to the problem of elasticity theory. Let $\tilde{T}$ be any tensor satisfying equations (5) and boundary condition (8) and let T_0 be the elastic stress tensor which solves our problem. Then T_0 is the projection of the tensor $\tilde{T}$ in the subspace $\mathfrak{H}_1$. Since $(\tilde{T}-T_0)\in\mathfrak{H}_2$, $(T_0, \tilde{T}-T_0)=0$ and

$$\|\tilde{T}\|^2=\|T_0\|^2+\|\tilde{T}-T_0\|^2\geq\|T_0\|^2. \tag{11}$$

Inequality (11) expresses Castigliano's principle: *of all the tensors which satisfy the equilibrium equations and the boundary condition at the free part of the boundary the actual elastic stress tensor is that which minimizes the potential energy of deformation to the body.*

We denote by $\mathbf{u}_0$ the displacement vector corresponding to the elastic stress tensor T_0. From formulae (14) § 26 and (4) § 11 we conclude that $\|\mathbf{u}_0\|^2=2W(\mathbf{u}_0)$, where $W(\mathbf{u}_0)$ is the potential energy of deformation corresponding to the displacement $\mathbf{u}_0$. Now formula (5) shows that $\|T\|=\|\mathbf{u}_0\|$; hence, as was shown above, there emerges the possibility of estimating the error of an approximate solution constructed by the Ritz method or by the method of orthogonal projections.

§ 55. Trefftz's method

Another method allowing us to estimate a lower bound for the required value of the energy functional is associated with the name of E. Trefftz [1].

The peculiarity of this method is the construction of some positive functional whose minimum is sought in the class of functions which satisfy the given differential equation but are not subjected to any boundary conditions. Trefftz himself discussed his method (but not its convergence) in its applications to the Dirichlet problem for the Laplace equation. We

also shall devote the present section to this case; in the following sections some generalizations of Trefftz's method will be discussed.

Let us require to find a function which is harmonic in the domain Ω and which satisfies the boundary condition

$$u|_S = f(P), \tag{1}$$

where $f(P)$ is a function which, for simplicity, we assume to be continuous on the boundary S. The function sought can be determined (see § 18) as the function as compared with any other function satisfying condition (1) which minimizes the integral

$$A(u) = \int_{\Omega} (\nabla u)^2 \mathrm{d}\Omega. \tag{2}$$

The Trefftz method consists of the following. We suppose that we are given a sequence of linearly independent functions φ_n which are harmonic in Ω and that this sequence is complete in the following sense: for any function φ which is harmonic and together with its first derivatives is quadratically summable in Ω it is possible to find for a specified $\varepsilon > 0$ a natural number n and constants $\alpha_1, \alpha_2, \ldots, \alpha_n$, such that

$$A\left(\varphi - \sum_{k=1}^{n} \alpha_k \varphi_k\right) = \int_{\Omega} \left\{\nabla\left(\varphi - \sum_{k=1}^{n} \alpha_k \varphi_k\right)\right\}^2 \mathrm{d}\Omega < \varepsilon.$$

We shall seek an approximate solution of our equation in the form

$$u_n = \sum_{k=1}^{n} a_k \varphi_k,$$

where n is an arbitrarily chosen number; we find the coefficients a_k from the condition

$$A(u - u_n) = \min, \tag{3}$$

where u is the required solution of our problem. By equating the derivatives $\dfrac{\partial A(u-u_n)}{\partial a_k}$ to zero, we readily arrive at the system of equations

$$A(u_n - u, \varphi_k) = 0, \qquad k = 1, 2, \ldots, n, \tag{4}$$

where for brevity we have used the notation

$$A(u, v)=\int_{\Omega}(\nabla u)\cdot(\nabla v)\mathrm{d}\Omega.$$

At first sight it would appear that system (4), as well as the coefficients a_k, depends upon the unknown function u. In fact this is not so. In order to be convinced of this we make use of Green's formula according to which

$$A(u_n-u,\ \varphi_k)=-\int_{\Omega}(u_n-u)\nabla^2\varphi_k\mathrm{d}\Omega+\int_{S}(u_n-u)\frac{\partial\varphi_k}{\partial n}\mathrm{d}S,$$

where n is the external normal to S. But $\nabla^2\varphi_k=0$, thus the first integral vanishes and we arrive at the system

$$\int_{S}(u_n-u)\frac{\partial\varphi_k}{\partial n}\mathrm{d}S=0,\quad k=1,\ 2,\ \ldots,\ n. \tag{5}$$

The value of the required function u is known on S and equals f. Replacing u by f and u_n by $\sum_{k=1}^{n}a_k\varphi_k$ we finally obtain the following system of linear algebraic equations

$$\sum_{j=1}^{n}a_j\int_{S}\varphi_j\frac{\partial\varphi_k}{\partial n}\mathrm{d}S=\int_{S}f\frac{\partial\varphi_k}{\partial n}\mathrm{d}S,\quad k=1,\ 2,\ \ldots,\ n. \tag{6}$$

We shall study system (6) closely. It is immediately clear that u_n does not change if u is changed by a constant, since this does not change the functional (3). Further, if u_n is an approximate solution obtained by the Trefftz method, then $u_n+\mathrm{const}$ will also be a solution. Bearing this in mind we shall assume that the mean values of the harmonic functions u and φ_k taken over the boundary S are equal to zero.

This is equivalent to the evaluation of an arbitrary constant in each such function. Now it is readily proved that the system (6) is soluble. Assuming the converse, we find that the homogeneous system

$$\sum_{j=1}^{n}a_j^{(0)}\int_{S}\varphi_j\frac{\partial\varphi_k}{\partial n}\mathrm{d}S=0,\quad k=1,\ 2,\ \ldots,\ n \tag{7}$$

has a non-trivial solution. We put $\sum_{j=1}^{n} a_j^{(0)} \varphi_j = v$. Then (7) assumes the form

$$\int_S v \frac{\partial \varphi_k}{\partial n} \, dS = 0, \quad k = 1, 2, \ldots, n.$$

Multiplying by $a_k^{(0)}$ and summing, we obtain

$$\int_S v \frac{\partial v}{\partial n} \, dS = 0.$$

By Green's formula

$$\int_S v \frac{\partial v}{\partial n} \, dS = \int_\Omega (\nabla v)^2 d\Omega - \int_\Omega v \nabla^2 v \, d\Omega.$$

The integral on the left equals zero; moreover $\nabla^2 v = 0$, since the function v is harmonic. Hence it follows that

$$\int_\Omega (\nabla v)^2 d\Omega = 0,$$

which is possible only if $\nabla v \equiv 0$ or $v \equiv \text{const}$. But since the mean value of $v|_S$ equals zero, $v = \sum_{j=1}^{n} a_j^{(0)} \varphi_j = 0$. Now from the linear independence of the functions φ_j it follows that $a_j^{(0)} = 0$, which is contrary to our supposition.

Once system (6) is solved, the approximate solution u_n can be constructed for any n. We now prove that

$$A(u_n) \leq A(u). \tag{8}$$

We put $u - u_n = w_n$. Then

$$A(u) = A(u_n + w_n) = A(u_n) + 2A(u_n, w_n) + A(w_n).$$

But

$$A(u_n, w_n) = A(u_n, u - u_n) = \sum_{k=1}^{n} a_k A(\varphi_k, u - u_n) = 0$$

by virtue of equations (4). Now

$$A(u) = A(u_n) + A(w_n),$$

whence inequality (8) immediately follows.

In conclusion we consider the question of the convergence of Trefftz's method. We restrict ourselves to the case of a plane domain Ω and we shall consider that its boundary S is smooth with continuous curvature. It can be proved the operator $\frac{\partial}{\partial n}$ is positive-bounded-below in the space $\overline{H}$ which consists of functions which are harmonic in Ω and representable through the Green function in terms of its limiting values. The latter are assumed to be quadratically summable and to satisfy the equation

$$\int_S u \mathrm{d}S = 0.$$

As $\frac{\partial}{\partial n}$ is a positive-bounded-below operator, there exists a positive constant γ for which the inequality (assuming v to be real)

$$\int_S v \frac{\partial v}{\partial n} \mathrm{d}S \geq \gamma^2 \int_S v^2 \mathrm{d}S, \quad v \in \overline{H} \tag{9}$$

is valid. Since the mean values of $u|_S$ and $\varphi_k|_S$ equal zero, these functions are elements of H. By the assumption of completeness for the sequence $\{\varphi_k\}$ it is possible for a specified $\varepsilon > 0$ to find a natural number N and constants $\alpha_1, \alpha_2, \ldots, \alpha_N$, such that

$$\Lambda\left(u - \sum_{k=1}^{N} \alpha_k \varphi_k\right) < \varepsilon.$$

A fortiori $\Lambda(u_N - u) < \varepsilon$, where u_N is an approximate solution constructed by Trefftz's method. For $n \geq N$, $\Lambda(u - u_n) \leq \Lambda(u - u_N) < \varepsilon$.

The function $u - u_n$ is harmonic and by Green's formula

$$\Lambda(u - u_n) = \int_S (u - u_n) \frac{\partial (u - u_n)}{\partial n} \mathrm{d}S.$$

By virtue of inequality (9)

$$\int_S (u - u_n)^2 \mathrm{d}S < \frac{\varepsilon}{\gamma^2}. \tag{10}$$

Let $G(P, Q)$ be Green's function in the domain Ω. Then

$$u(P)-u_n(P)=\int_S (u(Q)-u_n(Q))\frac{\partial G(P,Q)}{\partial n}\,dS. \qquad (11)$$

We suppose that P lies within some closed domain which lies entirely within Ω. Then as is known, the function $\frac{\partial G(P,Q)}{\partial n}$ and all its derivatives with respect to the coordinates of point P are bounded. Applying Cauchy's inequality to (11) and using the inequality (10), we find that $u-u_n\to 0$ *uniformly in any closed domain lying entirely within* Ω.

By differentiating (11), we may convince ourselves that the derivatives of u_n of any order converge uniformly to the corresponding derivative of u in any closed domain lying entirely within Ω.

We recall that we assume that the mean value both of u and of u_n are zero; without this the convergence $u_n\to u$, in general, does not occur. The statement about the convergence of the derivatives is valid independently of whether the mean values mentioned are zero or not.

In conclusion we note that in the construction of system (6) there is, in fact, no need to reduce the mean values of the functions u and φ_k to zero. In fact if we replace u by $u+c$ and φ_k by φ_k+c_k, where c and c_k are constants, then the coefficients and the independent term in (6) assume the following form:

$$\int_S \varphi_j\frac{\partial\varphi_k}{\partial n}dS+c_j\int_S\frac{\partial\varphi_k}{\partial n}dS; \qquad \int_S f\frac{\partial\varphi_k}{\partial n}dS+c\int_S\frac{\partial\varphi_k}{\partial n}dS.$$

But

$$\int_S\frac{\partial\varphi_k}{\partial n}dS=0,$$

since the function φ_k is harmonic. Hence it is clear that changing u and φ_k by constant amounts does not alter the system (6).

§ 56. The biharmonic equation. The method of anharmonic residue

A modification of the Trefftz method has been given by Z. Kh. Rafal'son [1, 2] for the case of a biharmonic equation. (We recall in this connexion the extremely interesting and considerably earlier work of S. Zaremba [3, 4] many of the results of which are close to the results of Z. Kh. Rafal'son.)

As was established in § 27, the integration of the biharmonic equation

$$\nabla^4 u = f \tag{1}$$

under the boundary conditions

$$u|_L = 0, \qquad \left.\frac{\partial u}{\partial n}\right|_L = 0 \tag{2}$$

is equivalent to the problem of finding the minimum of the functional

$$F(u) = (\nabla^4 u, u) - 2(u, f) = \int_S \{(\nabla^2 u)^2 - 2fu\} \mathrm{d}S \tag{3}$$

under the same conditions (2). The fact that this problem may be solved with the aid of the energy method was established in the same section.

The method of anharmonic residue is founded on the following theorem.

THEOREM 1. *Let $\tilde{u}(P)$ be any function which, together with its first and second derivatives, is continuous in S and which satisfies boundary conditions* (2) *on the boundary L of the domain S. Further, let $q(P)$ be any function which in S satisfies the equation*

$$\nabla^2 q = f, \tag{4}$$

and is continuous and continuously differentiable in $\bar{S} = S + L$ (assuming such a function exists).

Then

$$F(\tilde{u}) \geq -\int_S q^2 \mathrm{d}S.$$

The proof is extremely simple. We have

$$\int_S f\tilde{u}\,dS = \int_S \tilde{u}\nabla^2 q\,dS = \int_S q\nabla^2\tilde{u}\,dS + \int_L \left(\tilde{u}\frac{\partial q}{\partial n} - q\frac{\partial \tilde{u}}{\partial n}\right)ds = \int_S q\nabla^2\tilde{u}\,dS,$$

by virtue of conditions (2). Now

$$F(\tilde{u}) + \int_S q^2\,dS = \int_S \{(\nabla^2\tilde{u})^2 - 2\tilde{u}f + q^2\}\,dS$$

$$= \int_S \{(\nabla^2\tilde{u})^2 - 2q\nabla^2\tilde{u} + q^2\}\,dS = \int_S (\nabla^2\tilde{u} - q)^2\,dS \geq 0,$$

which was to be proved.

If we take $\tilde{u} = u$, where u is the solution of the biharmonic equation (1) which satisfies equation (2), then, as we have proved,

$$F(u) \geq -\int_S q^2\,dS. \tag{5}$$

Usually finding a solution of equation (4) presents no difficulty, and then the integral

$$-\int_S q^2\,dS = -\|q\|^2$$

gives the lower bound for the minimum of functional (5). It remains to to make this bound more precise. This is done in the following way. We take a sequence of functions $\{\varphi_k(P)\}$ which are *harmonic* in S and continuous and continuously differentiable in $\bar{S}$.

We shall consider them to be orthonormalized in S, so that

$$(\varphi_k, \varphi_m) = \int_S \varphi_k(P)\varphi_m(P)\,dS = \begin{cases} 0, & k \neq m \\ 1, & k = m. \end{cases}$$

This can always be done by applying the process of orthogonalization to the sequence adopted. We now put

$$q_n(P) = q(P) - \sum_{k=1}^{n} (q, \varphi_k)\varphi_k(P).$$

The function $q_n(P)$ satisfies all the conditions of theorem 1 and we now obtain as the lower bound of the quantity $F(u)$ the quantity $-\|q_n\|^2$ which, as is easily seen, equals

$$-\|q_n\|^2 = -\|q\|^2 + \sum_{k=1}^{n} (q, \varphi_k)^2 \tag{6}$$

and, consequently, is larger than $-\|q\|^2$. We now suppose that the sequence $\{\varphi_k(P)\}$ is complete for the set of functions which are harmonic in S and have a finite norm therein. Under well-known conditions it is then possible to prove that

$$F(u_0) = -\|q\|^2 + \sum_{k=1}^{\infty} (q, \varphi_k)^2, \tag{7}$$

where u_0 is the integral of equation (1) which satisfies (2).

The proof is based upon the following construction.

We consider the set G of functions which are harmonic and have a finite norm in S. A theorem due to H. Weyl [2] states that this set constitutes a subspace in $L_2(S)$. (A proof is also given by Z. Kh. Rafal'son [2]). The author [1] has extended the theorem to the more general case of a second order elliptic equation.

Subject to the well-known restrictions imposed upon the form of the domain S, it can be proved that functions from the subspace G permit, as precisely as desired, approximation in the mean by harmonic functions which are continuously differentiable in S; we shall assume that this is so here. This will be so, for instance, if the domain S is *stellate*, i.e. if there is a point within S such that any ray starting from it meets the boundary of the domain in not more than one point (see e.g. Z. Kh. Rafal'son [2]).

It is possible to indicate further sufficient criteria. Functions belonging to the orthogonal subspace $\overline{G} = L_2(S) \ominus G$ will be called *the anharmonic residues*. It follows from the definition itself that if $v(P)$ is a harmonic function quadratically summable in S, and $w(P)$ is some anharmonic residue, then

$$(v, w) = \int_S v(P) w(P) \mathrm{d}S = 0.$$

We put $\nabla^2 u_0 = w_0$. It is easily seen that w_0 is an anharmonic residue. In fact let $v(P)$ be a function which is harmonic in S and continuously differentiable in $\bar{S}$. Then

$$(w_0, v) = \int_S v \nabla^2 u_0 \mathrm{d}S$$

or, if we use Green's formula,

$$(w_0, v) = \int_S u_0 \nabla^2 v \mathrm{d}S + \int_L \left(v \frac{\partial u_0}{\partial n} - u_0 \frac{\partial v}{\partial n} \right) \mathrm{d}s = 0,$$

since $\nabla^2 v = 0$ and u_0 satisfies conditions (2). If $v(P)$ is not continuously differentiable in $\bar{S}$ then we approximate it by the continuously differentiable harmonic function $v_n(P)$; proceeding to the limit in the equation $(w_0, v_n) = 0$, we find that $(w_0, v) = 0$.

We have $\nabla^2 w_0 = \nabla^4 u_0 = f$. We put $q = v_0 + w_0$. Then $\nabla^2 v_0 = \nabla^2 q - \nabla^2 w_0$ and consequently $(w_0, v_0) = 0$. The function q is expanded into two orthogonal components $v_0 \in G$ and $w_0 \in \bar{G}$. In this case $\|q\|^2 = \|v_0\|^2 + \|w_0\|^2$, or

$$\|w_0\|^2 = \|q\|^2 - \|v_0\|^2. \tag{8}$$

From what has been said above it is clear that the orthonormal functions $\varphi_n(P)$ form a complete system in G. Hence

$$v_0(P) = \sum_{n=1}^{\infty} (v_0, \varphi_n) \varphi_n(P) \qquad \text{and} \qquad \|v_0\|^2 = \sum_{n=1}^{\infty} (v_0, \varphi_n)^2.$$

In which case

$$(v_0, \varphi_n) = (q, \varphi_n) - (w_0, \varphi_n) = (q, \varphi_n),$$

since w_0 and φ_n belong to mutually orthogonal subspaces G and $\bar{G}$ and therefore $(w_0, \varphi_n) = 0$. Now

$$\|v_0\|^2 = \sum_{n=1}^{\infty} (q, \varphi_n)^2.$$

Substituting this in (8) we obtain

$$\|w_0\|^2 = \|q\|^2 - \sum_{n=1}^{\infty} (q, \varphi_n)^2. \tag{9}$$

On the other hand

$$\|w_0\|^2=\|\nabla^2 u_0\|^2=\int_S(\nabla^2 u_0)^2 dS.$$

Comparing this with formula (15) § 27, we find that $\|w_0\|^2 = \|u_0\|^2 = -F_{\min}(u)$ and formula (7) is proved. By retaining a finite number of terms in series (7) we obtain a quantity less than $F_{\min}(u)$ and as close to it as required if the number of terms taken is sufficiently large.

From the calculations given above there readily emerges the formula

$$w_0(P)=\nabla^2 u_0=q(P)=-\sum_{n=1}^{\infty}(q,\varphi_n)\varphi_n(P). \tag{10}$$

Knowing $w_0(P)$ we can establish the function $u_0(P)$ according to the formula

$$u(P)=\frac{1}{2\pi}\int_S\int w_0(Q)\ln r\,dS_Q, \tag{11}$$

where r is the distance between the points P and Q. The derivation of formula (11) is given in the paper by Z.Kh. Rafal'son [2].

§ 57. Generalization of Trefftz's method

Trefftz's method was generalized for a wider class of boundary value problems in the works of M.Sh. Birman [3], [5], [6]. We suppose that it is required to integrate in the domain Ω the linear differential equation

$$Lu=f(P) \tag{1}$$

under specified boundary conditions. For simplicity we shall consider these to be homogeneous:

$$G_1u|_S=0,\quad G_2u|_S=0,\quad \ldots,\quad G_ru|_S=0. \tag{2}$$

Let the operator A which coincides with the differential operator L for functions which satisfy conditions (2) be positive-

bounded-below. Then as we know, the function $u_0(P)$ which satisfies equations (1) and (2) realizes the minimum of the functional

$$F(u)=(Au,\ u)-2(u,\ f), \tag{3}$$

for simplicity we consider only real functions)

$$F_{\min}u=-\|u_0\|^2=-(Au_0,\ u_0). \tag{4}$$

The generalized Trefftz's method is constructed, in general outline, in the following manner: a new functional $\varphi(u)$ is constructed which is defined for functions which need only satisfy differential equation (1). The satisfaction of any kind of boundary condition is not necessary. We require of the functional $\varphi(u)$ that it assume a minimum value for the function $u_0(P)$ which satisfies equations (1) and (2) and that

$$\varphi_{\min}(u)=\varphi(u_0)=\|u_0\|^2. \tag{5}$$

We note that for solutions of equation (1) $\varphi(u)$ is necessarily positive—this follows from formula (5).

Now let $u(P)$ be any solution of equation (1). From formula (5) it then follows that $\varphi(u)\geq\|u_0\|^2$ and, consequently, $-\varphi(u)\leq F_{\min}(u)$. Thus $-\varphi(u)$ can be taken as the number δ in inequalities (4) and (5) § 48. Further, let $u_n(P)$, $n=1, 2, \ldots$, be a sequence of the solutions of equation (1) which is minimizing for functional $\varphi(u)$, so that

$$\lim_{n\to\infty}\varphi(u_n)=\|u\|^2.$$

Then $-\varphi(u_n)\to F_{\min}(u)$ and we obtain a sequence of numbers $\delta_n=-\varphi(u_n)$ which for sufficiently large n differ by as little as required from $d=F_{\min}(u)$.

In each concrete boundary value problem we are thus concerned with the selection of an appropriate functional $\varphi(u)$.

In §§ 58 and 59 we shall briefly discuss the construction due to M. Sh. Birman for the functionals $\varphi(u)$ for the more important boundary value problems connected with the Poisson equation and the equation for the bending of a plate; as we shall see the functional $\varphi(u)$ is not constructed uniquely,

since for each boundary value problem there is an infinite set of "Trefftz methods". We note further that we usually succeed in proving the convergence of the minimizing sequence to the exact solution of the problem; thus the Trefftz method also proves convenient as a further method for the approximate solution of boundary value problems.

We note that in the paper by M. Sh. Birman cited above functionals of the generalized Trefftz method are also constructed for various more complex problems of the theory of thin plates.

§ 58. Application to Poisson's equation

We consider the problem of integrating Poisson's equation

$$-\nabla^2 u = f(P) \tag{1}$$

under the mixed boundary conditions ($S_1 + S_2 = S$)

$$u|_{S_1} = 0, \tag{2}$$

$$\left[\frac{\partial u}{\partial n} + \sigma u\right]_{S_2} = 0, \qquad \sigma > 0. \tag{3}$$

It is easily shown by integrating by parts and using the boundary conditions, that these are equivalent to

$$\|u\|^2 = (-\nabla^2 u,\ u) = \int_{\Omega} (\nabla u)^2 \mathrm{d}\Omega + \int_{S_2} \sigma u^2 \mathrm{d}S. \tag{4}$$

In order to simplify the discussion we assume that the solution of our problem is a function which is continuously differentiable in the closed domain $\overline{\Omega} = \Omega + S$.

We choose some non-negative function $\tilde{\sigma}$ which is defined over the entire surface S and satisfies the inequality

$$\tilde{\sigma} < \sigma \quad \text{on} \quad S_2. \tag{5}$$

We now put

$$\varphi(u) = \int_{\Omega} (\nabla u)^2 \mathrm{d}\Omega + \int_{S} \tilde{\sigma} u^2 \mathrm{d}S + \int_{S_2} \frac{1}{\sigma - \tilde{\sigma}} \left(\frac{\partial u}{\partial n} + \tilde{\sigma} u\right)^2 \mathrm{d}S. \tag{6}$$

The corresponding "bilinear" functional has the form

$$\varphi(u, v)=\int_{\Omega}(\nabla u)\cdot(\nabla v)\mathrm{d}\Omega+\int_{S}\tilde{\sigma}uv\mathrm{d}S$$

$$+\int_{S_2}\frac{1}{\sigma-\tilde{\sigma}}\left(\frac{\partial u}{\partial n}+\tilde{\sigma}u\right)\left(\frac{\partial v}{\partial n}+\tilde{\sigma}v\right)\mathrm{d}S; \tag{7}$$

$$\varphi(u+v)=\varphi(u)+2\varphi(u, v)+\varphi(v). \tag{8}$$

Obviously $\varphi(u)\geq 0$, always. We note two further important properties of the functional $\varphi(u)$:

(a) If the function $u(P)$ satisfies boundary conditions (2) and (3), then $\varphi(u)=\|u\|^2$. In fact, in this case on S_2, $\dfrac{\partial u}{\partial n}=-\sigma u$. Substituting this in (6) we readily find

$$\varphi(u)=\int_{\Omega}(\nabla u)^2\mathrm{d}\Omega+\int_{S_2}\sigma u^2\,\mathrm{d}S+\int_{S_1}\tilde{\sigma}u^2\mathrm{d}S.$$

By virtue of condition (2) the last integral vanishes and comparison with formula (4) gives $\varphi(u)=\|u\|^2$.

(b) If $u(P)$ satisfies conditions (2) and (3) and $v(P)$ is an arbitrary harmonic function which is continuously differentiable in $\overline{\Omega}$ then $\varphi(u, v)=0$. In fact, taking the first integral in (7) by parts and bearing in mind the boundary conditions which the function $u(P)$ satisfies, we obtain

$$\varphi(u, v)=-\int_{\Omega}u\nabla^2 v\mathrm{d}\Omega+\int_{S}u\frac{\partial v}{\partial n}\,\mathrm{d}S+\int_{S}\tilde{\sigma}uv\mathrm{d}S$$

$$+\int_{S_2}\frac{1}{\sigma-\tilde{\sigma}}\left(\frac{\partial u}{\partial n}+\tilde{\sigma}u\right)\left(\frac{\partial v}{\partial n}+\tilde{\sigma}v\right)\mathrm{d}S$$

$$=\int_{S_2}\frac{1}{\sigma-\tilde{\sigma}}\left(\frac{\partial u}{\partial n}+\sigma u\right)\left(\frac{\partial v}{\partial n}+\tilde{\sigma}v\right)\mathrm{d}S=0.$$

THEOREM 1. *Of all the solutions of equation* (1) *which are continuously differentiable in* $\overline{\Omega}$, *the solution* $u_0(P)$ *which satis-*

fies boundary conditions (2) *and* (3) *minimizes the functional* $\varphi(u)$; *and*

$$\varphi_{\min}(u)=\varphi(u_0)=\|u_0\|^2. \tag{9}$$

Let $u(P)$ be a solution of (1) which is continuously differentiable in $\overline{\Omega}$. We put $u=u_0+v$, so that $\nabla^2 v=0$, i.e. the function v is harmonic. By property (b) $\varphi(u_0, v)=0$. Now from formula (8) it follows that for $v\not\equiv 0$, i.e. for $u\not\equiv u_0$,

$$\varphi(u)=\varphi(u_0+v)=\varphi(u_0)+\varphi(v)>\varphi(u_0). \tag{10}$$

But this also implies that $\varphi_{\min}(u)=\varphi(u_0)$. By property (a) $\varphi(u_0)=\|u_0\|^2$ since u_0 satisfies boundary conditions (2) and (3). The theorem is proved; it is simultaneously established that the functional $\varphi(u)$ satisfies all the requirements of § 57.

Now let $u_n(P)$, $n=1, 2, \ldots$ be a sequence of the solutions of equation (2) which is minimizing for the functional $\varphi(u)$, so that

$$\lim_{n\to\infty}\varphi(u_n)=\varphi(u_0).$$

By formula (10) we have

$$\varphi(u_n)=\varphi(u_n-u_0)+\varphi(u_0).$$

Hence it immediately follows that

$$\lim_{n\to\infty}\varphi(u_n-u_0)=0.$$

On the basis of formula (6) we conclude that $\nabla u_n\to\nabla u_0$ in the mean over Ω. If further $\tilde{\sigma}>0$, then $u_n\to u_0$ in the mean over the surface S_1 and $\dfrac{\partial u_n}{\partial n}\to\dfrac{\partial u_0}{\partial n}$ in the mean over S_2.

We shall dwell upon the two most important particular cases. 1. Let S_2 be absent, so that $u=0$ on the entire surface S, and our problem is Dirichlet's problem for equation (1). In this case

$$\varphi(u)=\int_{\Omega}(\nabla u)^2 d\Omega+\int_S\tilde{\sigma}u^2 dS, \tag{*}$$

where $\tilde{\sigma}$ is an arbitrary function which is non-negative on S.

If u_n, $n=1, 2, \ldots$ is a minimizing sequence for the functional (*) then, as was indicated above, $\nabla u_n \to \nabla u_0$ in the mean in Ω, and if $\tilde{\sigma} > 0$, then $u_n \to u_0$ in the mean over S. Using the frequently cited "imbedding theorems" due to S. L. Sobolev, it can be proved that $u_n \to u_0$ in the mean in Ω; since the u_n are solutions of equation (1) then, as is known, it follows that $u_n \to u_0$ uniformly together with all derivatives in any closed domain lying entirely within Ω.

It is possible to put $\tilde{\sigma} \equiv 0$ in (*) and this will then exactly correspond to the Trefftz method itself. In this case if u_n is the minimizing sequence then it is only possible to assert that $\nabla u_n \to \nabla u_0$ and it is not possible to assert the convergence of the functions themselves. We met with this circumstance in § 55.

2. We consider the Neumann problem

$$\left.\frac{\partial u}{\partial n}\right|_S = 0. \tag{11}$$

As we know (§ 22), this problem is soluble if

$$\int_\Omega f \, d\Omega = 0,$$

and the solution is unique if in addition we require that

$$\int_\Omega u \, d\Omega = 0.$$

In this case it is possible to put

$$\varphi(u) = \int_\Omega (\nabla u)^2 d\Omega - \alpha \int_S u^2 dS + \frac{1}{\alpha} \int_S \left(\frac{\partial u}{\partial n} - \alpha u \right)^2 dS, \tag{12}$$

where α is a sufficiently small positive constant. It can be proved (see M. Sh. Birman [6]) that functional (12) satisfies all the requirements formulated in § 57.

In the Neumann problem the generalized Trefftz method has the advantage over the method of orthogonal projections that in seeking the minimum of functional (12) there is no need to subject the function u to any boundary conditions.

In the Dirichlet problem this advantage is lost, since in the method of orthogonal projections the vectors w are not subject to any boundary conditions.

The generalized Trefftz method is readily extended to elliptic second-order equations with variable coefficients and also to the case of non-homogeneous boundary conditions.

§ 59. Generalization of Trefftz's method to the problem of a freely supported plate

Passing on to problems concerned with the bending of thin plates connected with biharmonic equations we shall restrict ourselves to the two simplest cases of a rigidly fixed and of a freely supported edge, and for each of them we shall indicate one (in our opinion, the simplest) form of the functional of Trefftz's method. For the case of a rigidly fixed edge such a form was presented in § 56; in the present section we shall consider a plate with a freely supported edge.

We precede the construction of the Trefftz method functional with the derivation of an auxiliary identity. We put formula (6) § 27 in the form

$$2\int_{\Omega}\int\left[\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2-\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}\right]dxdy=\int_{L}\left[\frac{\partial w}{\partial y}d\left(\frac{\partial w}{\partial x}\right)-\frac{\partial w}{\partial x}d\left(\frac{\partial w}{\partial y}\right)\right]$$

$$=\int_{L}\left(\frac{\partial w}{\partial y}\frac{\partial^2 w}{\partial x^2}dx+\frac{\partial w}{\partial y}\frac{\partial^2 w}{\partial x\partial y}dy-\frac{\partial w}{\partial x}\frac{\partial^2 w}{\partial x\partial y}dx-\frac{\partial w}{\partial x}\frac{\partial^2 w}{\partial y^2}dy\right).$$

It is easily seen that the expression in the line integral is invariant under a translation of the origin and under a rotation of the cartesian coordinate axes. Bearing this in mind we replace the x and y axes by the external normal n and the tangent s for each point on L. Then obviously $dn=0$ along the edge, and we obtain

$$2\int_{S}\int\left[\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2-\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}\right]dxdy=\int_{L}\left(\frac{\partial w}{\partial s}w_{ns}-\frac{\partial w}{\partial n}w_{ss}\right)ds. \quad (1)$$

In formula (1) the quantities w_{ns} and w_{ss} are the second derivatives with respect to the directions of the tangent and the normal evaluated on the assumption that the directions of the tangent and normal do not change; more specifically

$$w_{ns}=\frac{\partial^2 w}{\partial x^2}\cos(n,\,x)\cos(s,\,x)+\frac{\partial^2 w}{\partial x\partial y}[\cos(n,\,x)\cos(s,\,y)$$

$$+\cos(n,\,y)\cos(s,\,x)]+\frac{\partial^2 w}{\partial y^2}\cos(n,\,y)\cos(s,\,y);$$

$$w_{ss}=\frac{\partial^2 w}{\partial x^2}\cos^2(s,\,x)+2\frac{\partial^2 w}{\partial x\partial y}\cos(s,\,x)\cos(s,\,y)$$

$$+\frac{\partial^2 w}{\partial y^2}\cos^2(s,\,y).$$

An argument similar to that presented in § 27 shows that

$$w_{ns}=\frac{\partial}{\partial s}\left(\frac{\partial w}{\partial n}\right)-\frac{1}{\varrho}\frac{\partial w}{\partial s};\quad w_{ss}=\frac{\partial^2 w}{\partial s^2}+\frac{1}{\varrho}\frac{\partial w}{\partial n},$$

where ϱ is the radius of curvature of the edge. Substituting the latter expression in (1), we obtain

$$2\int_S\int\left[\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2-\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}\right]dS$$

$$=-\int_L\left[\frac{1}{\varrho}\left(\frac{\partial w}{\partial n}\right)^2+\frac{1}{\varrho}\left(\frac{\partial w}{\partial s}\right)^2+\frac{\partial w}{\partial n}\frac{\partial^2 w}{\partial s^2}-\frac{\partial w}{\partial s}\frac{\partial}{\partial s}\left(\frac{\partial w}{\partial n}\right)\right]ds.$$

Since the curve L is closed and the function w and its derivatives are assumed to be single-valued, integration by parts gives

$$\int_L\frac{\partial w}{\partial s}\frac{\partial}{\partial s}\left(\frac{\partial w}{\partial n}\right)ds=-\int_L\frac{\partial w}{\partial n}\frac{\partial^2 w}{\partial s^2}\,ds,$$

which leads to the formula required

$$2\int_S\int\left[\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2-\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}\right]dS$$

$$=-\int_L\left[2\frac{\partial w}{\partial n}\frac{\partial^2 w}{\partial s^2}+\frac{1}{\varrho}\left(\frac{\partial w}{\partial n}\right)^2+\frac{1}{\varrho}\left(\frac{\partial w}{\partial s}\right)^2\right]ds. \qquad (2)$$

Replacing w in (2) firstly by $u+v$ and then by $u-v$ and subtracting, we obtain another useful formula:

$$\int_S\int\left(2\frac{\partial^2 u}{\partial x\partial y}\frac{\partial^2 v}{\partial x\partial y}-\frac{\partial^2 u}{\partial x^2}\frac{\partial^2 v}{\partial x^2}-\frac{\partial^2 u}{\partial y^2}\frac{\partial^2 v}{\partial x^2}\right)\mathrm{d}S$$

$$=-\int_L\left(\frac{\partial u}{\partial n}\frac{\partial^2 v}{\partial s^2}+\frac{\partial v}{\partial n}\frac{\partial^2 u}{\partial s^2}+\frac{1}{\varrho}\frac{\partial u}{\partial n}\frac{\partial v}{\partial n}+\frac{1}{\varrho}\frac{\partial u}{\partial s}\frac{\partial v}{\partial s}\right)\mathrm{d}s. \quad (3)$$

In the problem of a freely supported plate it is necessary to construct the functional of the generalized Trefftz method in a different manner according as to whether the edge of the plate is everywhere convex or whether it contains non-convex parts. If the edge L is convex, then we can put

$$\varphi(w)=\int_S\int\left[\left(\frac{\partial^2 w}{\partial x^2}\right)^2+2\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2+\left(\frac{\partial^2 w}{\partial y^2}\right)^2\right]\mathrm{d}S$$

$$+\frac{1}{\sigma}\int_L\varrho\left[\nabla^2 w-\left(\frac{\partial^2 w}{\partial s^2}+\frac{1}{\varrho}\frac{\partial w}{\partial n}\right)\right]^2\mathrm{d}s; \quad (4)$$

here σ is Poisson's ratio which we consider positive.

We shall prove that functional (4) possesses all the properties enumerated in § 58.

(a) For any function w we have $\varphi(w)\geq 0$, since the edge L is convex and consequently $\varrho\geq 0$.

(b) We construct the bilinear functional

$$\varphi(u,v)=\int_S\int\left(\frac{\partial^2 u}{\partial x^2}\frac{\partial^2 v}{\partial x^2}+2\frac{\partial^2 u}{\partial x\partial y}\frac{\partial^2 v}{\partial x\partial y}+\frac{\partial^2 u}{\partial y^2}\frac{\partial^2 v}{\partial y^2}\right)\mathrm{d}S$$

$$+\frac{1}{\sigma}\int_L\varrho\left[\nabla^2 u-\left(\frac{\partial^2 u}{\partial s^2}+\frac{1}{\varrho}\frac{\partial u}{\partial n}\right)\right]\left[\nabla^2 v-\left(\frac{\partial^2 v}{\partial s^2}+\frac{1}{\varrho}\frac{\partial v}{\partial n}\right)\right]\mathrm{d}s. \quad (5)$$

We shall prove that $\varphi(u,v)=0$ if u satisfies the conditions for a freely supported edge

$$u|_L=0, \qquad \left[\nabla^2 u-\frac{1-\sigma}{\varrho}\frac{\partial u}{\partial n}\right]_L=0, \quad (6)$$

and v satisfies the biharmonic equation

$$\nabla^4 v = 0. \tag{7}$$

For the proof we note that by virtue of conditions (6) on the edge L

$$\nabla^2 u - \left(\frac{\partial^2 u}{\partial s^2} + \frac{1}{\varrho}\frac{\partial u}{\partial n}\right) = \nabla^2 u - \frac{1}{\varrho}\frac{\partial u}{\partial n} = -\frac{\sigma}{\varrho}\frac{\partial u}{\partial n},$$

which allows us to simplify the line integral in (5). Further

$$\int_S\int\left(\frac{\partial^2 u}{\partial x^2}\frac{\partial^2 v}{\partial x^2} + 2\frac{\partial^2 u}{\partial x \partial y}\frac{\partial^2 v}{\partial x \partial y} + \frac{\partial^2 u}{\partial y^2}\frac{\partial^2 v}{\partial y^2}\right) \mathrm{d}S$$

$$= \int_S\int \nabla^2 u \nabla^2 v \mathrm{d}S + \int_S\int\left(2\frac{\partial^2 u}{\partial x \partial y}\frac{\partial^2 v}{\partial x \partial y} - \frac{\partial^2 u}{\partial x^2}\frac{\partial^2 v}{\partial y^2}\right.$$

$$\left. - \frac{\partial^2 u}{\partial y^2}\frac{\partial^2 v}{\partial x^2}\right) \mathrm{d}S,$$

which by virtue of formulae (3) and (6) assumes the form

$$\int_S\int \nabla^2 u \nabla^2 v \mathrm{d}S - \int_L \left(\frac{\partial^2 v}{\partial s^2} + \frac{1}{\varrho}\frac{\partial v}{\partial n}\right)\frac{\partial u}{\partial n}\,\mathrm{d}s.$$

Now

$$\varphi(u, v) = \int_S\int \nabla^2 u \nabla^2 v \mathrm{d}S - \int_L \nabla^2 v \frac{\partial u}{\partial n}\,\mathrm{d}s.$$

Replacing v by $\nabla^2 v$ in Green's formula [formula (10) § 2] we find

$$\int_S\int \nabla^2 u \nabla^2 v \mathrm{d}S = \int_S\int u \nabla^4 v \mathrm{d}S + \int_L\left(\frac{\partial u}{\partial n}\nabla^2 v - u\frac{\partial \nabla^2 v}{\partial n}\right)\mathrm{d}s.$$

Hence

$$\varphi(u, v) = \int_S\int u \nabla^4 v \mathrm{d}S - \int_L u \frac{\partial(\nabla^2 v)}{\partial n}\,\mathrm{d}s,$$

which vanishes by virtue of equations (6) and (7).

(c) Let w be an arbitrary solution of the equation for the bending of the plate

$$D\nabla^4 w = q,$$

and let w_0 be the solution of this equation which satisfies the conditions of free support. The function $v = w - w_0$ satisfies equation (7); as proved $\varphi(w_0, v) = 0$. Now we have

$$\varphi(w) = \varphi(w_0 + v) = \varphi(w_0) + 2\varphi(w_0, v) + \varphi(v) = \varphi(w_0) + \varphi(v).$$

Hence it is clear that

$$\varphi_{\min}(w) = \varphi(w_0).$$

(d) We now evaluate $\varphi(w_0)$. Using condition (6) which the function satisfies, we find

$$\varphi(w_0) = \int_S \int \left[\left(\frac{\partial^2 w_0}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 w_0}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 w_0}{\partial y^2} \right)^2 \right] dS + \sigma \int_L \frac{1}{\varrho} \left(\frac{\partial w_0}{\partial n} \right)^2 ds.$$

The same condition (6) allows us to represent identity (2) in the form

$$\int_L \frac{1}{\varrho} \left(\frac{\partial w_0}{\partial n} \right)^2 ds = 2 \int_S \int \left[\frac{\partial^2 w_0}{\partial x^2} \frac{\partial^2 w_0}{\partial y^2} - \left(\frac{\partial^2 w_0}{\partial x \partial y} \right)^2 \right] dS.$$

Substituting this in $\varphi(w_0)$, we obtain

$$\varphi(w_0) = \int_S \int \left[\left(\frac{\partial^2 w_0}{\partial x^2} \right)^2 + 2\sigma \frac{\partial^2 w_0}{\partial x^2} \frac{\partial^2 w_0}{\partial y^2} + \left(\frac{\partial^2 w_0}{\partial y^2} \right)^2 + 2(1-\sigma) \left(\frac{\partial^2 w_0}{\partial x \partial y} \right)^2 \right] dS;$$

by formula (21) § 27, $\varphi(w_0) = \|w_0\|^2$

If the edge L contains non-convex parts (where $\varrho < 0$), then functional (4) ceases to be positive definite and becomes

unsuitable for our purposes. In this case we can put (cf. M. Sh. Birman [6])

$$\varphi(w)=\int_S\int\left[\left(\frac{\partial^2 w}{\partial x^2}\right)^2+2\sigma\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}+\left(\frac{\partial^2 w}{\partial y^2}\right)^2+2(1-\sigma)\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2\right]\mathrm{d}S$$

$$+\int_L\left\{\beta w^2-\alpha\left(\frac{\partial w}{\partial n}\right)^2+\frac{1}{\alpha}\left[\nabla^2 w-(1-\sigma)\left(\frac{\partial^2 w}{\partial s^2}+\frac{1}{\varrho}\frac{\partial w}{\partial n}\right)\right]^2\right\}\mathrm{d}s, \tag{8}$$

where α and β are positive constants and the constant α is sufficiently small.

§ 60. The method of M. G. Slobodyansky

In obtaining a lower bound for the minimal functional it is convenient in some cases to use a simple procedure due to M. G. Slobodyansky [8]. This procedure, which is a generalization of the Friedrichs transformation mentioned above, consists of the following.

Let A be a symmetric, positive-bounded-below operator,

$$(Au,\, u)\geq\gamma^2\|u\|^2, \tag{1}$$

and let u_0 be the solution of the equation

$$Au=f. \tag{2}$$

We assume that operator A can be represented in the form of the sum

$$A=\sum_{k=1}^{p}A_k, \tag{3}$$

where each operator A_k is positive definite and symmetric. Let U denote a selection of elements $u_1, u_2, \ldots, u_p$ such that

u_k belongs to the field of definition of the operator A_k and satisfies the equation

$$\sum_{k=1}^{p} A_k u_k = f. \tag{4}$$

We construct the functional

$$\Psi(U) = \sum_{k=1}^{p} (A_k u_k, u_k). \tag{5}$$

We have

THEOREM 1. *The functional $\Psi(U)$ is a minimum for $u_k = u_0$, $k = 1, 2, \ldots, p$; and*

$$\min \Psi(U) = \|u_0\|^2. \tag{6}$$

We put $\eta_k = u_k - u_0$. Then

$$\sum_{k=1}^{p} A_k \eta_k = \sum_{k=1}^{p} A_k u_k - \sum_{k=1}^{p} A_k u_0 = f - Au_0 = 0. \tag{7}$$

Now

$$\Psi(U) = \sum_{k=1}^{p} (A_k(u_0 + \eta_k), u_0 + \eta_k)$$

$$= (Au_0, u_0) + \sum_{k=1}^{p} (A_k \eta_k, u_0) + \sum_{k=1}^{p} (A_k u_0, \eta_k) + \sum_{k=1}^{p} (A_k \eta_k, \eta_k).$$

By virtue of equation (7)

$$\sum_{k=1}^{p} (A_k \eta_k, u_0) = 0; \quad \sum_{k=1}^{p} (A_k u_0, \eta_k) = \sum_{k=1}^{p} (u_0, A_k \eta_k) = 0.$$

and we obtain

$$\Psi(U) = \|u_0\|^2 + \sum_{k=1}^{p} (A_k \eta_k, \eta_k).$$

Since the operator A_k is positive definite, $\Psi(U) \geq \|u_0\|^2$; the equality sign is possible only if $\eta_k = 0$, $k = 1, 2, \ldots, p$, i.e. if $u_k = u_0$, $k = 1, 2, \ldots, p$.

In the most general case it is possible to denote a positive-bounded-below operator A by the form $A_1 + A_2$, if we take any constant c such that $0 < c < \gamma$ and put $A_1 u = Au - cu$,

$A_2u=cu$. In this case $\Psi(U)=(Au_1-cu_1,\ u_1)+c(u_2,\ u_2)$; the elements u_1 and u_2 are connected by the equation

$$Au_1-cu_1+cu_2=f.$$

Hence

$$\Psi(U)=(Au_1-cu_1,\ u_1)+\frac{1}{c}(f-Au_1+cu_1,\ f-Au_1+cu_1); \quad (8)$$

here u_1 is an arbitrary element from the field of definition of operator A.

We further note some particular cases. If

$$Au=-\sum_{j,\,k=1}^{m}\frac{\partial}{\partial x_j}\left(A_{jk}(P)\frac{\partial u}{\partial x_k}\right)+C(P)u,$$

where $C(P)$ is a strictly positive function, and if the boundary conditions are such that the operator

$$A_1u=-\sum_{j,\,k=1}^{m}\frac{\partial}{\partial x_j}\left(A_{jk}(P)\frac{\partial u}{\partial x_j}\right)$$

is positive definite for the set of functions which satisfy these conditions, then it is possible to put $A_2u=C(P)u$ and $A=A_1+A_2$. The function u_1 can be chosen arbitrarily, but such that it satisfies the boundary conditions of the problem. u_2 is then determined from the equation

$$-\sum_{j,\,k=1}^{m}\frac{\partial}{\partial x_j}\left(A_{jk}\frac{\partial u_1}{\partial x_k}\right)+C(P)u_2=f(P);$$

in which case

$$\Psi(U)=-\int_{\Omega}u_1\sum_{i,\,k=1}^{m}\frac{\partial}{\partial x_j}\left(A_{jk}\frac{\partial u_1}{\partial x_k}\right)\mathrm{d}\Omega+\int_{\Omega}Cu_2^2\mathrm{d}\Omega.$$

In particular, if the boundary condition has the form $u|_S=0$,

then the function u_1 also satisfies this condition. Integrating by parts we reduce $\Psi(U)$ to the form

$$\Psi(U)=\int_{\Omega}\left\{\sum_{j,k=1}^{m}A_{jk}\frac{\partial u_1}{\partial x_j}\frac{\partial u_1}{\partial x_k}+Cu_2^2\right\}d\Omega.$$

We further consider the equation

$$Au=\sum_{k=0}^{m}(-1)^k\frac{d^k}{dx^k}\left[p_k(x)\frac{d^k u}{dx^k}\right]=f(x)$$

under the boundary conditions

$$u^{(k)}(a)=u^{(k)}(b)=0, \qquad k=0,\ 1,\ 2,\ \ldots,\ m-1.$$

If the coefficients are such that $p_m(x)\geq\tilde{p}>0$, $p_k^{(x)}\geq 0$, then it is possible to put

$$A_k u=(-1)^k\frac{d^k}{dx^k}\left[p_k(x)\frac{d^k u}{dx^k}\right], \quad k=0,\ 1,\ 2,\ \ldots,\ m;$$

$$u^{(j)}(a)=u^{(j)}(b)=0, \qquad j\leq k-1.$$

In our case $U=(u_0, u_1, u_2, \ldots, u_m)$; the functions u_k are connected by the equation

$$\sum_{k=0}^{m}(-1)^k\frac{d^k}{dx^k}\left[p_k(x)\frac{du}{dx^k}\right]=f(x).$$

The functional $\Psi(U)$ has the form:

$$\Psi(U)=\int_a^b\sum_{k=0}^{m}p_k(x)\left(\frac{d^k u_k}{dx^k}\right)^2dx.$$

§ 61. Upper and lower bounds for functionals

It is sometimes important to find, not the solution of the equation

$$Au=f, \tag{1}$$

itself, but only the value of some functional of the form (u, g) (where g is some given function) which is connected with this solution. From a number of treatments of this problem we select the work of M. G. Slobodyansky [6] a brief exposition of which will be given in the present section (see also the works [3], [4], [9] of this author); at the end of the section we shall also refer to the interesting but more difficult work of T. Kato [1].

Let A be a positive-bounded-below, symmetric operator and let it be required to evaluate the scalar product (u_0, g), where g is a given function (or in the more general case, a given element of the Hilbert space considered) and u_0 is the solution of equation (1). The simplest way of solving this problem is as follows: we construct an approximate solution u_n of equation (1) and we put approximately

$$(u_0, g) = (u_n, g). \tag{2}$$

The error of this approximate equation is readily estimated using Cauchy's inequality:

$$|(u_0, g) - (u_n, g)| = |(u_0 - u_n, g)| \leq \|g\| \|u_0 - u_n\|. \tag{3}$$

If u_n was constructed, for example, by the energy method, then the quantity $\|u_0 - u_n\|$ can be estimated by using the method of orthogonal projections. Equation (2) can be made more accurate.

Let v_0 satisfy the equation

$$Av = g \tag{4}$$

and let v_n be its approximate solution. We put

$$Au_n = f_n, \quad Av_n = g_n, \quad b_n = (u_n, g) + (f - f_n, v_n). \tag{5}$$

We put approximately

$$(u_0, g) = b_n \tag{6}$$

and we estimate the error of this equation. The identity

$$(u_0, g) - b_n = (u_0 - u_n, g - g_n)$$

is readily verified. The error required is

$$|(u_0-u_n,\ g-g_n)| = |(u_0-u_n,\ A(v_0-v_n)| = |[u_0-u_n,\ v_0-v_n]|,$$

and by Cauchy's inequality

$$|(u_0,\ g)-b_n| \leq \|u_0-u_n\|\,\|v_0-v_n\|. \tag{7}$$

The right-hand side can be estimated by the methods which have been established in the present chapter if u_n and v_n are constructed by the energy method or by the method of orthogonal projections. (In the work of M. G. Slobodyansky the construction of a formula which is more accurate than formula (6) is indicated.)

We shall further present, without dwelling on the derivation, some results of T. Kato [1]. Let $A=T^*T$, where we impose the conditions of § 52 upon the operators A, T^*, T. As before, let u_0 satisfy equation (1). Further et u and v be arbitrary elements from the field of definition of operator T and let u' and v' be arbitrary solutions of the equations $T^*u'=f$, $T^*v'=g$. We put

$$\alpha=(u,\ g)+(f,\ v)-(Tu,\ Tv),\ \beta=(u',\ v').$$

Then the error of the approximate equation

$$(u_0,\ g)=\frac{1}{2}(\alpha+\beta) \tag{8}$$

does not exceed the value

$$\frac{1}{2}\|Tu-u'\|\,\|Tv-v'\|. \tag{9}$$

§ 62. Upper and lower bounds of eigenvalues

1. As we know the Ritz method gives upper bounds for the eigenvalues; the problem in this section is to indicate methods for obtaining lower bounds for the eigenvalues. It is not possible to consider this problem as entirely solved. In fact, at the present time it is doubtful whether it is possible to indicate a procedure (or group of procedures) which would allow us effectively to construct for any positive-bounded-below operator, approximate lower bounds for the eigenvalues with any degree of precision.

In this section we shall indicate two particular procedures each of which has a fairly restricted field of application. The first of these is connected with the method of integral equations and is due to the author [7]; the second procedure which was given by I. V. Svirsky [1] is based on the minimax property of eigenvalues.

2. In many cases we can construct the so called Green's function $G(P, Q)$ of a given symmetric and positive-bounded-below operator A. Then the eigenvalues and eigenfunctions of operator A can be found as the characteristic values and eigenfunctions of the integral equation

$$u(P)-\lambda\int_{\Omega} G(P, Q)u(Q)\mathrm{d}\Omega_Q=0. \tag{1}$$

For simplicity we shall restrict ourselves to the case most frequently encountered where the "double" integral

$$\int_{\Omega}\int_{\Omega} G^2(P, Q)\mathrm{d}\Omega_P\mathrm{d}\Omega_Q$$

has a finite value. We note that from the symmetry of operator A there follows the symmetry of the Green function corresponding to it.

Considering $G(P, Q)$ as the kernel of an integral equation, we shall denote by $G_m(P, Q)$, $m=1, 2, \ldots$ the integrated kernels corresponding to it which are defined by the well-known recurrence relations:

$$G_1(P, Q)=G(P, Q), \quad G_m(P, Q)=\int_{\Omega} G(P, R)G_{m-1}(R, Q)\mathrm{d}\Omega_R.$$

We introduce the so called *traces* of the integrated kernels

$$a_m=\int_{\Omega} G_m(P, P)\mathrm{d}\Omega.$$

For the traces with even indices the formula

$$a_{2m}=\int_{\Omega}\int_{\Omega} G_m^2(P, Q)\mathrm{d}\Omega_P\mathrm{d}\Omega_Q, \tag{*}$$

is valid, which allows us to evaluate the trace using half the number of integrations. If the traces a_{2m} have been evaluated

then it is simple to find the lowest eigenvalue for equation (1). Thus the traces a_m are related to the eigenvalues $\lambda_1, \lambda_2, \ldots$ by the relation

$$a_m = \sum_{n=1}^{\infty} \frac{1}{\lambda_n^m}, \qquad m \geq 2. \tag{2}$$

Let the eigenvalues be arranged in order of size.
Replacing m by $2m$ in the last equation and dropping all the terms in the series except the first, we obtain

$$\lambda_1 \frac{1}{\sqrt[2m]{a_{2m}}}. \tag{3}$$

Thus the approximate equation

$$\lambda_1 \approx \frac{1}{\sqrt[2m]{(a_{2m})}} \tag{4}$$

gives lower bound for λ_1. It can be proved that

$$\lambda_1 = \lim_{m \to \infty} \frac{1}{\sqrt[2m]{a_{2m}}},$$

so that for n sufficiently large the approximate equation (4) will be as precise as required. If the multiplicity p of the eigenvalue λ is known then, together with (3), it is possible to obtain the more precise inequality

$$\lambda_1 \geq \sqrt[2m]{\left(\frac{p}{a_{2m}}\right)} \tag{5}$$

and the correspondingly more precise approximate equation for a lower bound

$$\lambda_1 \approx \sqrt[2m]{\left(\frac{p}{a_{2m}}\right)}. \tag{6}$$

We note that from series (2) it is also possible to obtain an upper bound for λ_1:

$$\lambda_1 \approx \sqrt{\left(\frac{a_{2m}}{a_{2m+2}}\right)}, \tag{7}$$

which in the limit becomes the exact formula

$$\lambda_1 = \lim_{m\to\infty} \sqrt{\left(\frac{a_{2m}}{a_{2m+2}}\right)}.$$

It is also possible to construct approximate formulae for the successive eigenvalues of equation (1). Thus, if the eigenvalues λ_1 and λ_2 are both simple, then

$$\lambda_2 \approx \frac{1}{\lambda_1} \sqrt[2m]{\left(\frac{2}{a_{2m}^2 - a_{4m}}\right)}. \tag{8}$$

If the exact value (or an upper bound) of λ_1 is known, then formula (8) gives a lower bound λ_2; as $m\to\infty$ we obtain the exact formula

$$\lambda_2 = \frac{1}{\lambda_1} \lim_{m\to\infty} \frac{1}{\sqrt[2m]{(a_{2m}^2 - a_{4m})}}.$$

The detailed derivation of the formulae cited here and also some numerical examples are given in the author's book [7], §§ 15 and 17. The procedure formulated above often gives good results if A is a differential operator in one independent variable, since then in many cases the evaluation of the Green function and its integration required little effort.

3. Let A and B be symmetric, positive-bounded-below operators and let $B \leq A$ (cf. § 40). Then as we know, the eigenvalues of operator B are smaller than the corresponding eigenvalues of operator A. If operator B is constructed and its eigenvalues are readily found then we have at our disposal approximate (although, in general, crude) lower bounds for the eigenvalues of operator A. Examples in which it is possible to construct such a "lesser" operator are readily pointed out. Thus, for example, we denote by A the operator $-\nabla^2$ specified for functions in the domain Ω and vanishing on its boundary, and by B, the same operator $-\nabla^2$ but specified for functions defined in a more extensive domain Ω_1 which vanish on its boundary. We shall prove that $B < A$. To this end we extend the functions from H_A

by putting them equal to zero in $\Omega_1 - \Omega$. Then they can be considered as functions from H_B. Thus H_B is wider than H_A. At the same time if $u \in H_A$ then

$$\|u\|_A^2 = \int_{\Omega} (\nabla u)^2 \mathrm{d}\Omega = \int_{\Omega_1} (\nabla u)^2 \mathrm{d}\Omega_1 = \|u\|_B^2.$$

By the definition given in § 40, $B < A$ and the eigenvalues of operator B are smaller than the corresponding eigenvalues of operator A. If, for example, we take a parallelepiped as Ω_1 then the eigenvalues of operator B are readily found.

We now give another, simpler example. Let

$$Au = -\frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}u}{\mathrm{d}x}\right) + r(x)u,$$

and the functions $u \in D_A$ be defined by the boundary conditions $u(0) = u(1) = 0$. Let $p(x) \geq p_0 > 0$ and $r(x) \geq r_0 \geq 0$ where p_0 and r_0 are constants. It is possible to take as the operator with the required property the operator B defined by

$$Bu = -p_0 \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} + r_0 u$$

with the same boundary conditions, $u(0) = u(1) = 0$. The eigenvalues of operator B can be found in elementary fashion; they equal $r_0 + p_0 n^2 \pi^2$, $n = 1, 2, \ldots$ and, since obviously $B < A$, then $\lambda_n > r_0 + p_0 n^2 \pi^2$, where λ_n is the nth eigenvalue of operator A.

I.V. Svirsky procedure mentioned at the beginning of this section allows us to construct operators A_n such that $B \leq A_n \leq A$. If in this case the lowest eigenvalue of the operator A_n is strictly greater than the lowest eigenvalue of operator B, then we obtain a more precise lower bound for the lowest eigenvalue of the given operator A.

We put $A - B = C$. The operator C is non-negative; for simplicity (although this is not necessary), we assume that it is positive-bounded-below. We select linearly independent functions (or, more generally, elements of the given Hilbert space) $f_1, f_2, \ldots, f_n$ which belong to the field of definition of

operator C. By applying the process of orthogonalization, if necessary, we can ensure that these functions are orthonormalized in the energy of operator C, so that

$$[f_k, f_m]_C = (Cf_k, f_m) = \begin{cases} 0, & k \neq m, \\ 1, & k = m. \end{cases} \tag{9}$$

We shall use the notation

$$Cf_k = g_k, \qquad k = 1, 2, \ldots, n \tag{10}$$

and put

$$C_n u = \sum_{k=1}^{n} [u, f_k]_C g_k = \sum_{k=1}^{n} (Cu, f_k) g_k \tag{11}$$

and

$$A_n = B + C_n. \tag{12}$$

We shall prove that C_n is a symmetric operator satisfying the inequality $0 \leq C_n \leq C$; hence it will follow that $B \leq A_n \leq A$. We have

$$(C_n u, v) = \sum_{k=1}^{n} [u, f_k]_C (g_k, v) = \sum_{k=1}^{n} (Cu, f_k)(Cf_k, g),$$

or, since operator C is symmetric,

$$(C_n u, v) = \sum_{k=1}^{n} (Cu, f_k)(f_k, Cv).$$

We obtain the same result by evaluating the scalar product $(u, C_n v)$; hence the symmetry of operator C is proved.

In order to prove the inequality $0 \leq C_n \leq C$ we evaluate the quantity

$$(C_n u, u) = \sum_{k=1}^{n} [u, f_k](Cf_k, u) = \sum_{k=1}^{n} [u, f_k][f_k, u] = \sum_{k=1}^{n} |[u, f_k]|^2 \geq 0,$$

(where we have omitted the symbol C from the energy product).

Hence $C_n \geq 0$. On the other hand, by Bessel's inequality

$$\sum_{k=1}^{n} |[u, f_k]_C|^2 \leq \|u\|_C^2 = (Cu, u).$$

4. It is easily explained what the nature of the functions f_k must be in order that the lowest eigenvalues $\mu_1^{(n)}$ of operator A_n be strictly greater than the lowest eigenvalue μ_1 of operator B. Let φ_1 be the normalized eigenfunction of operator B corresponding to the eigenvalue μ_1. We have

$$\mu_1 = \inf_{\|u\|=1} (Bu, u) = (B\varphi_1, \varphi_1),$$

$$\mu_1^{(n)} = \inf_{\|u\|=1} (A_n u, u) = \inf_{\|u\|=1} \left\{ (Bu, u) + \sum_{k=1}^{n} |[u, f_k]_C|^2 \right\}. \quad (13)$$

The first term on the right assumes a minimum value equal to μ_1 for $u=\varphi_1$. If the function φ_1 is orthogonal in the energy of operator C to all the functions f_k, then, obviously, $\mu_1^{(n)}=\mu_1$, and in this case the method of I. V. Svirsky gives nothing. Now let at least one product $[\varphi_1, f_k]\neq 0$ and let the minimum in (13) be achieved for $u=\tilde{\varphi}$:

$$\mu_1^{(n)} = (B\tilde{\varphi}, \tilde{\varphi}) + \sum_{k=1}^{n} |[\tilde{\varphi}, f_k]_C|^2. \quad (14)$$

If $\tilde{\varphi}\neq\varphi$ then $(B\tilde{\varphi}, \tilde{\varphi}) > \mu_1$; if $\tilde{\varphi}=\varphi_1$, then the sum in (14) is positive. In both cases $\mu_1^{(n)} > \mu_1$. Thus, passing from operator B to operator A improves the approximation below to the lowest eigenvalue of operator A if and only if at least one of the functions f_k is not orthogonal (in the energy of operator C) to one of the eigenfunctions of operator B which corresponds to the lowest eigenvalue of this operator.

§ 63. Estimate of the error due to a perturbation of the equation

It is often advantageous to replace the given equation

$$Au = f \quad (1)$$

by a close but simpler equation, which we shall write in the form

$$Bu = f. \quad (2)$$

The question arises of how much such a substitution affects the solutions. In other words, how can we estimate the degree of closeness of the solutions of equations (1) and (2) if we can somehow judge the degree of closeness of the operators A and B.

In the present section we shall formulate some results contained in our paper [9] (see also the works [10] and [17] of the author); we omit the proofs since they lie beyond the scope of this book.

We shall assume that the operators A and B are symmetric and positive-bounded-below (it would be possible to restrict ourselves to the requirement of positive definiteness only for the operators A and B). We shall suppose that the spaces H_A and H_B (§ 46) corresponding to these operators are coincident, in other words, that every function whose energy with respect to operator A is finite also has finite energy with respect to operator B, and vice versa. It can be proved that there are then positive constants α_1 and α_2 such that for any function $u \in H_A$ (or, what is the same thing, $u \in H_B$) the inequality

$$\alpha_1 \|u\|_B^2 \leq \|u\|_A^2 \leq \alpha_2 \|u\|_B^2 \tag{3}$$

is valid. When the operators A and B are given the determination of the constants α_1 and α_2 presents no particular difficulty and we shall consider these constants to be known.

If u_0 and u_1 are solutions of equations (1) and (2), so that $Au_0 = f$ and $Bu_1 = f$ then the inequality

$$\|u_1 - u_0\|_B \leq \eta \|u_1\|_B, \tag{4}$$

holds, in which the constant η is defined by the formula

$$\eta = \max\left(\frac{|\alpha_1 - 1|}{\alpha_1}; \frac{|\alpha_2 - 1|}{\alpha_2}\right). \tag{5}$$

Formulae (4) and (5) solve the problem posed at the beginning of the section.

As an example we consider the Neumann problem for the elliptic equation

$$Au=-\sum_{j,k=1}^{m}\frac{\partial}{\partial x_j}\left(A_{jk}\frac{\partial u}{\partial x_k}\right)=f(P). \tag{6}$$

This problem consists (cf. § 24) in seeking a function $u(P)$ which is defined in some domain Ω in the coordinate space $(x_1, x_2, x_3, \ldots, x_m)$; this function must satisfy equation (6) within the domain Ω, and the boundary condition

$$\left[\sum_{j,k=1}^{m}A_{jk}\frac{\partial u}{\partial x_j}\cos(n, x_k)\right]_S=0 \tag{7}$$

(n is the external normal) on the boundary of the domain Ω. As usual, we consider the domain Ω finite and the boundary piecewise smooth. We shall consider the coefficients A_{jk} to bounded and equation (6) to be non-degenerate. Then it is possible to prove the existence of two positive constants μ_0 and μ_1 such that, for any numbers $t_1, t_2, \ldots, t_m$ and any position of the point P within or on the boundary of the domain Ω the inequality

$$\mu_0\sum_{j=1}^{m}t_j^2\leq\sum_{j,k=1}^{m}A_{jk}(P)t_jt_k\leq\mu_1\sum_{j=1}^{m}t_j^2 \tag{8}$$

is satisfied. We recall that for the solubility of the Neumann problem it is necessary and sufficient that

$$\int_{\Omega}f(P)\mathrm{d}\Omega=0. \tag{9}$$

The solution will be unique if we subject it to the condition

$$\int_{\Omega}u(P)\mathrm{d}\Omega=0. \tag{10}$$

We shall assume both of these conditions to be fulfilled.

We introduce the real Hilbert space H as the subspace of $L_2(\Omega)$ defined by equation (10). In this space the operator is positive-bounded-below for the set of functions which satisfy boundary condition (7). In this case

$$\|u\|_A^2=\int_{\Omega}\sum_{i,k=1}^{m}A_{jk}\frac{\partial u}{\partial x}\frac{\partial u}{\partial x_k}\mathrm{d}\Omega. \tag{11}$$

Condition (7) is natural, therefore the space H_A (in other words, the set of functions with finite energy with respect to operator A) consists of all the functions which give a finite value to integral (11); it follows from inequality (8) that H_A can be considered as the set of functions which give a finite value to the simpler integral

$$\int_{\Omega}\sum_{j=1}^{m}\left(\frac{\partial u}{\partial x_j}\right)^2\mathrm{d}\Omega. \tag{12}$$

We shall solve the Neumann problem for the elliptic equation

$$-\sum_{j,k=1}^{m}\frac{\partial}{\partial x_j}\left(B_{jk}\frac{\partial v}{\partial x_k}\right)=f(P). \tag{13}$$

In the same domain Ω; the boundary condition has the form

$$\left[\sum_{j,k=1}^{m}B_{jk}\frac{\partial v}{\partial x_j}\cos\,(n,\,x_k)\right]_S=0. \tag{14}$$

The space H_B consists of functions which give finite value to integral (12), i.e. of the same functions as the space H_A.

We construct an equation in the unknown λ

$$\begin{vmatrix} B_{11}-\lambda A_{11} & B_{12}-\lambda A_{12} & \dots & B_{1n}-\lambda A_{1n} \\ B_{21}-\lambda A_{21} & B_{22}-\lambda A_{22} & \dots & B_{2n}-\lambda A_{2n} \\ \dots & \dots & \dots & \dots \\ B_{n_1}-\lambda A_{n_1} & B_{n_2}-\lambda A_{n_2} & \dots & B_{nn}-\lambda A_{nn} \end{vmatrix}=0. \tag{15}$$

All the roots of this equation are positive (cf. § 32). Let $\lambda'(P)$ be the least; and $\lambda''(P)$ the greatest root of equation (15).

Then, as is known from linear algebra, (see, for example, R. Courant and D. Hilbert [1], Ch. 1)

$$\lambda'(P)\sum_{j,k=1}^{m}B_{jk}(P)t_jt_k\leq\sum_{j,k=1}^{m}A_{jk}(P)t_jt_k\leq\lambda''(P)\sum_{j,k=1}^{m}B_{jk}t_jt_k. \tag{16}$$

Inequality (16) can be obtained by applying the reasoning of § 34 to the linear transformations which are defined by the matrices of the coefficients $A_{jk}(P)$ and $B_{jk}(P)$, where the t_j, are any real numbers. We now use the notation

$$\alpha_1=\inf_{P\in\Omega}\lambda'(P);\qquad \alpha_2=\sup_{P\in\Omega}\lambda''(P).$$

We replace λ' by α_1 and λ'' by α_2 in inequality (16), thus strengthening the inequality, and we put $t_j=\dfrac{\partial u}{\partial x_j}$. Integrating the equation obtained over Ω, we obtain

$$\alpha_1\|u\|_B^2\leq\|u_A^2\|\leq\alpha_2\|u\|_B^2$$

Denoting by u and v the solutions of the Neumann problem for equations (16) and (12), we have $\|u-v\|\leq\eta\|v\|$ where η is defined by formula (5). The last inequality shows, in particular, that if $A_{jk}(P)\to B_{jk}(P)$ uniformly in the closed domain, then both the function $u(P)$ itself and its first partial derivatives converge in the mean in Ω to the function $v(P)$ and to its first partial derivatives.

CHAPTER VIII

NUMERICAL EXAMPLES

THE AIM of the present chapter is to indicate by way of examples the applications of the methods of the preceding chapters, to elucidate the computational difficulties thus encountered and to point out ways of surmounting them. One of the basic difficulties connected with the application of direct methods is the construction of a complete system of coordinate functions. In view of the extreme importance of this question we shall start the present chapter with a section which is theoretical in nature. In succeeding sections we shall consider numerical examples of boundary value problems and eigenvalue problems, we shall also estimate the error of the constructed approximation wherever it is of importance.

§ 64. On coordinate functions

1. The application of the Ritz method requires the initial selection of a system of coordinate functions which is complete in the sense of energy convergence. The proof of the completeness of the system chosen can often be carried out by using the following theorem.

THEOREM 1. *Let A be a positive-bounded-below, symmetric operator and let φ_n, $n=1, 2, \ldots$ be functions from its field of definition. If the set of functions $\{A\varphi_n\}$ is complete in the sense of convergence in the mean, then the set of functions $\{\varphi_n\}$ is complete in the sense of energy convergence.*

Proof. Let $v(P)$ be an arbitrary function from the field of definition of operator A. Since the system $\{A\varphi_n\}$ is complete in the sense of convergence in the mean it is possible to find a positive integer n and constants $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that

$$\left\| Av - \sum_{k=1}^{n} \alpha_k A\varphi_k \right\| = \left\| A\left(v - \sum_{k=1}^{n} \alpha_k \varphi_k \right) \right\| < \frac{1}{2}\gamma\varepsilon, \qquad (1)$$

where ε is an arbitrary positive number and γ is the constant in formula (7) § 6.

We now estimate the quantity $\left| v - \sum_{k=1}^{n} \alpha_k \varphi_k \right|$. Denoting for brevity, $\sum_{k=1}^{n} \alpha_k \varphi_k$ by v_n, we have

$$|v - v_n|^2 = (A(v - v_n), v - v_n).$$

Using Cauchy's inequality and then inequality (1), we obtain

$$|v - v_n|^2 \leq \|A(v - v_n)\| \, \|v - v_n\| < \frac{1}{2}\varepsilon\gamma \|v - v_n\|. \qquad (2)$$

By inequality (2) § 8, $\|v - v_n\| \leq \frac{1}{\gamma}|v - v_n|$. Substituting this in inequality (2) of the present section and dividing by $|v - v_n|$, we obtain

$$|v - v_n| < \frac{1}{2}\varepsilon. \qquad (3)$$

Now let $u(P)$ be an arbitrary function with finite energy. By definition (§ 16) there exists a function $v \in D_A$ such that $|u - v| < \frac{1}{2}\varepsilon$. For the function v_n we choose the numbers $n, \alpha_1, \alpha_2, \ldots, \alpha_n$ such that inequality (3) is satisfied. Now by the triangle inequality

$$\left| u - \sum_{k=1}^{n} \alpha_k \varphi_k \right| = |u - v_n| \leq |u - v| + |v - v_n| < \varepsilon,$$

which was to be proved.

NOTE: Theorem 1 can obviously be extended to the case of an arbitrary Hilbert space. In this case it is formulated thus:

Let A be an operator which is positive-bounded-below and symmetric in some Hilbert space H, and let $\varphi_n \in D_A$, $n=1, 2, \ldots$. [If the Hilbert space is complex, it is then unnecessary to stipulate the symmetry of operator A.]

If the sequence $\{A\varphi_n\}$ is complete in H then the sequence $\{\varphi_n\}$ is complete in H_A.

Example. Let Ω be the rectangle $0 \leq x \leq a$, $0 \leq y \leq b$ and let $Au = -\nabla^2 u$, with u vanishing on the boundary of the rectangle. Let

$$\varphi_{m,n}(x, y) = \sin\frac{m\pi x}{a} \sin\frac{n\pi y}{b} \qquad m, n = 1, 2 \ldots; \tag{4}$$

then

$$A\varphi_{m,n} = -\nabla^2\varphi_{m,n} = \pi^2\left(\frac{m^2}{a^2} + \frac{n^2}{b^2}\right)\sin\frac{m\pi x}{a}\sin\frac{n\pi y}{b}.$$

From the theory of double Fourier series it is known that the system of functions

$$\sin\frac{m\pi x}{a}\sin\frac{n\pi y}{a}; \qquad m, n = 1, 2, \ldots$$

is complete in the sense of convergence in the mean; since the multiplication of each function by a non-zero constant does not violate the completeness of the system, the system $A\varphi_{n,m}$ is complete in the sense of convergence in the mean in the rectangular domain Ω; it follows from theorem 1 that the system (4) is complete in the same domain in the sense of energy convergence. In particular, in the given case, this implies the following: let $u(x\ y)$ be any function which is continuous and continuously differentiable within the rectangle $0 \leq x \leq a$, $0 \leq y \leq b$ and vanishes at its edges; for any positive number ε (no matter how small) it is possible to find positive integers m and n and constants α_{kl}, $k=1, 2, \ldots, m$; $l=1, 2, \ldots, n$ such that the inequality

$$\int_0^a\int_0^b\left\{\left(\frac{\partial u}{\partial x} - \frac{\pi}{a}\sum_{k=1}^{m}\sum_{l=1}^{n} k\alpha_{kl}\cos\frac{k\pi x}{a}\sin\frac{l\pi y}{b}\right)^2\right.$$

$$\left.+\left(\frac{\partial u}{\partial y} - \frac{\pi}{b}\sum_{k=1}^{m}\sum_{l=1}^{n} l\alpha_{kl}\sin\frac{k\pi x}{a}\cos\frac{l\pi y}{b}\right)^2\right\}dx\,dy < \varepsilon^2$$

is satisfied. We note that the completeness of system (4) in the space H_A can also be proved on the basis of theorem 3 § 31, since system (4) is the system of eigenfunctions of the Laplacian operator corresponding to the boundary condition $u|_S = 0$ over the rectangle.

We leave it to the reader to prove that the system of functions

$$(x-a)^m(b-x)^m x^k, \qquad k=0,\ 1,\ 2,\ \ldots$$

is complete in the energy of the operator

$$(-1)^m \frac{d^{2m}u}{dx^{2m}}$$

under the boundary conditions

$$u^{(k)}(a)=u^{(k)}(b)=0, \qquad k=0,\ 1,\ \ldots,\ m-1.$$

2. We now indicate a method of constructing coordinate functions which although given here for a particular problem nevertheless proves useful in a wide class of cases.

We consider the Dirichlet problem in some domain Ω for Poisson's equation under the condition that the required function vanishes on the boundary S of the domain Ω. In this case functions with finite energy vanish on S and

$$\|u\|^2_{-\nabla^2}=\int_\Omega\int\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]dxdy.$$

To be explicit, we assume that we are concerned with functions of the two independent variables x and y. A system of coordinate functions which is complete in energy can be constructed in the following way.

Let $\omega(x, y)$ be a function which vanishes on the boundary S and is positive at points within the domain Ω; we further assume that this function is continuous in $\bar{\Omega}$ and that its first derivatives are continuous and bounded within Ω. Then the system of functions

$$\omega(x, y)x^k y^l; \qquad k,\ l=1,\ 2,\ \ldots \tag{5}$$

is complete in energy within Ω. The proof of this important assertion was given by V. I. Krylov ([1] pp. 294–295). The construction of the function $\omega(x, y)$ usually encounters no difficulty. Thus, for example, for the rectangle referred to above, we can take $\omega=xy(a-x)(b-y)$. Nor does the transition to functions of a larger number of independent variables present any difficulty. The system of coordinate functions (5) or its analogue for the case of a larger number of independent variables can be used in solving not only the Poisson equation, but also more complex problems. Consider the

integration of a non-degenerate equation of elliptic type (A, B, C, depend on x and y):

$$Lu=-\frac{\partial}{\partial x}\left(A\frac{\partial u}{\partial x}+B\frac{\partial u}{\partial y}\right)-\frac{\partial}{\partial y}\left(B\frac{\partial u}{\partial x}+C\frac{\partial u}{\partial y}\right)=f(x,\ y)$$

in a given domain Ω under the boundary condition $u|_S=0$. In this case

$$\|u\|_L^2=\int_\Omega\int\left[A\left(\frac{\partial u}{\partial x}\right)^2+2B\frac{\partial u}{\partial x}\frac{\partial u}{\partial y}+C\left(\frac{\partial u}{\partial y}\right)^2\right]dxdy.$$

By assuming that the given equation is non-degenerate, there automatically follows the existence of a positive constant μ_0 such that

$$A\left(\frac{\partial u}{\partial x}\right)^2+2B\frac{\partial u}{\partial x}\frac{\partial u}{\partial y}+C\left(\frac{\partial u}{\partial y}\right)^2\geq\mu_0\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]. \quad (6)$$

We further suppose that the coefficients A, B, C are bounded. Then, obviously there exists a positive constant μ_1 such that

$$A\left(\frac{\partial u}{\partial x}\right)^2+2B\frac{\partial u}{\partial x}\frac{\partial u}{\partial y}+C\left(\frac{\partial u}{\partial y}\right)^2\leq\mu_1\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]. \quad (7)$$

From inequalities (6) and (7) it follows that

$$\sqrt{(\mu_0)}\|u\|_{-\nabla^2}\leq\|u\|_L\leq\sqrt{(\mu_1)}\|u\|_{-\nabla^2}.$$

Hence we readily see that any system, in particular system (5), which is complete in the energy of the operator $-\nabla^2$ is also complete in the energy of operator L.

3. For the biharmonic operator under the boundary conditions $u|_S=0$, $\left.\frac{\partial u}{\partial n}\right|_S=0\ldots.$, the system of functions

$$\omega^2x^ky^l;\quad k,l=0,1,2,\ldots,$$

where $\omega(x,y)$ is the function described in subsection **2**, is complete in energy.

In general for the polyharmonic operator $(-1)^p\nabla^{2p}$ under the boundary conditions

$$u|_S=\left.\frac{\partial u}{\partial n}\right|_S=\ldots=\left.\frac{\partial^{p-1}u}{\partial n^{p-1}}\right|_S=0$$

the system of functions

$$\omega^p x^k y^l; \qquad k, l = 0, 1, 2, \ldots$$

is complete in energy.

4. Suppose, for example, that we are concerned with the Dirichlet problem for Poisson's equation. The fact that system (5) is complete in energy means the following: if $u(x, y)$ is an arbitrary function with finite energy and ε is an arbitrary positive number then there exists a polynomial $P(x, y)$ such that

$$\int_\Omega \int \left[\left(\frac{\partial u}{\partial x} - \frac{\partial(\omega P)}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} - \frac{\partial(\omega P)}{\partial y} \right)^2 \right] dx dy < \varepsilon^2. \qquad (8)$$

In some cases it is possible to establish a stronger result. Suppose that $\omega(x, y)$ is a function which is k times continuously differentiable in $\overline{\Omega}$ and that $\frac{\partial \omega}{\partial n} \neq 0$ on S. Further let the function $u(x, y)$ vanish on S and be continuously differentiable k times in $\overline{\Omega}$. Then there exists a sequence of polynomials $P_n(x, y)$ of degree less than or equal to n in both x and y for which the following estimate holds:

$$|D^r u - D^r(\omega P_n)| < \frac{C}{n^{k-r-1}}, \quad r = 0, 1, \ldots, k-1,$$

where C is a constant and D^r denotes any derivative of order r. [Cf. I. Yu. Kharrik [1] in which a rather stronger assertion is proved.]

It is possible to extract from inequality (8), an estimate of the rapidity of the convergence of the Ritz method. Let the required function $u_0(x, y)$ be continuously differentiable k times within $\overline{\Omega}$. Putting $r = 1$ in (8) we find that there exists a polynomial $P_n(x, y)$ such that

$$\left| \frac{\partial u_0}{\partial x} - \frac{\partial(\omega P_n)}{\partial x} \right| \leq \frac{C}{n^{k-2}}, \quad \left| \frac{\partial u_0}{\partial y} - \frac{\partial(\omega P_n)}{\partial y} \right| \leq \frac{C}{n^{k-2}}.$$

Hence

$$\|u_0-\omega P_n\|=\left\{\int_\Omega\left[\left(\frac{\partial u_0}{\partial x}-\frac{\partial(\omega P_n)}{\partial x}\right)^2+\left(\frac{\partial u_0}{\partial y}-\frac{\partial(\omega P_n)}{\partial y}\right)^2\right]d\Omega\right\}^{\frac{1}{2}}\leq\frac{C_1}{n^{k-2}},$$

where C_1 is some new constant. We now take the set of functions (5) as coordinate functions and we construct an approximate solution by the Ritz method. This solution has the form $u_n=\omega Q_n$, where Q_n is a polynomial whose degree we can choose equal to n. This is equivalent to using some known number of coordinate functions. The Ritz method, for given coordinate functions, leads to the least value of the norm of the energy $\|u-u_n\|$. Hence we can conclude that $\|u_0-\omega Q_n\|\leq\|u_n-\omega P_n\|$ and consequently,

$$\|u_0-\omega Q_n\|\leq\frac{C_1}{n^{k-2}}. \tag{9}$$

5. If u_n is an approximate solution constructed by the Ritz method, then, as we have seen, $u_n\to u$ in the sense of convergence in the mean, and also in the sense of energy convergence. (If the equation is considered in an arbitrary Hilbert space H, then $u_n\to u$ in the sense of convergence in space H and also in the sense of convergence in space H_A.) It is not possible in addition to this to assert that in the general case $Au_n\to f$. In other words, in the general case the assertion that the Ritz approximate solution satisfies the given equation is invalid.

In some cases, however, such an assertion is valid. Firstly it is trivially valid if A is a bounded operator. Now let A be an unbounded, positive-bounded-below, symmetric operator. We shall call operator B *similar* to operator A if: (1) B is a positive-bounded-below, symmetric operator; (2) the fields of definition of the operators A and B coincide. We have

THEOREM 2. *Let an operator B which is similar to operator A have a discrete spectrum. Let the eigenfunctions of operator*

B be taken as the coordinate functions and suppose that with their aid we construct a Ritz approximate solution u_n for the equation $Au=f$. Then $Au_n \to f$ in the sense of convergence in the mean (cf. the author's paper [18]).

Examples 1. Under the boundary conditions of the Dirichlet problem (S is the boundary of the domain)

$$u|_S = 0$$

the operators $Bu=-\nabla^2 u$ and

$$Au=-\sum_{j,k=1}^{m}\frac{\partial}{\partial x_j}\left(A_{jk}\frac{\partial u}{\partial x_k}\right)+C(P)u$$

are similar, if the operator A is non-degenerate. Therefore in solving the Dirichlet problem for the equation

$$-\sum_{j,k=1}^{m}\frac{\partial}{\partial x_j}\left(A_{jk}\frac{\partial u}{\partial x_k}\right)+C(P)u=f(P)$$

it is expedient to take as the coordinate functions the eigenfunctions of the Laplacian operator for the domain Ω which satisfy the boundary condition $u|_S=0$.

2. Under the boundary condition for the Neumann problem

$$\left.\frac{\partial u}{\partial n}\right|_S = 0$$

the positive-bounded-below operators $-\nabla^2 u+u$ and $-\nabla^2 u+C(P)u$, where $C(P)>0$ are similar. In solving the Neumann problem for the equation $-\nabla^2 u+C(P)u=f(P)$ it is expedient to take as the coordinate functions the eigenfunctions (satisfying the boundary condition $\left.\frac{\partial u}{\partial n}\right|_S=0$) of the operator $-\nabla^2 u+u$, or, what is the same thing, of the Laplacian operator for the given domain Ω.

§ 65. The torsion of a rod of rectangular cross-section

1. Solution by the Ritz method. We shall consider in some detail this extremely simple problem because we know a comparatively simple exact solution for it, with which we can readily compare our approximations. This problem reduces, as is known, to the integration of the Poisson equation:

$$-\nabla^2\psi=2G\theta$$

(G is the shear modulus, θ is the angle of torsion per unit length of rod) in the rectangle $-a \leq x \leq a$, $-b \leq y \leq b$, under the boundary conditions $\psi(\pm a, y) = \psi(x, \pm b) = 0$. In order to simplify the calculation we put $\psi = 2G\theta u$, then

$$-\nabla^2 u = 1, \tag{1}$$

$$u(\pm a, y) = u(x, \pm b) = 0. \tag{2}$$

In § 12 we gave the solution of this problem in the form of an expansion into a series which is orthogonal in energy. We now apply the Ritz method to the same problem taking polynomials for the coordinate functions.

From symmetry it is clear that the function $u(x, y)$ is even both in x and in y. Polynomials possessing this property and vanishing on the edges of the rectangle, i.e. on the straight lines $x = \pm a$, $y = \pm b$ have the form

$$(x^2 - a^2)(y^2 - b^2)(a_1 + a_2 x^2 + a_3 y^2 + \ldots).$$

We restrict ourselves to three terms and we put as our approximate solution

$$u \approx u_3 = (x^2 - a^2)(y^2 - b^2)(a_1 + a_2 x^2 + a_3 y^2). \tag{3}$$

By performing the necessary calculations we find the Ritz system of equations for the unknowns a_1, a_2, a_3:

$$\left.\begin{aligned}
&\frac{128}{45} a^3 b^3 (a^2 + b^2) a_1 + \frac{128}{45} a^5 b^3 \left(\frac{a^2}{7} + \frac{b^2}{5}\right) a_2 \\
&\qquad + \frac{128}{45} a^3 b^5 \left(\frac{a^2}{5} + \frac{b^2}{7}\right) a_3 = \frac{16 a^3 b^3}{9}, \\
&\frac{128}{45} a^5 b^3 \left(\frac{a^2}{7} + \frac{b^2}{5}\right) a_1 + \frac{128}{45 \times 7} a^5 b^5 \left(\frac{11}{5} b^2 + \frac{1}{3} a^2\right) a_2 \\
&\qquad + \frac{128}{45 \times 35} a^5 b^5 (a^2 + b^2) a_3 = \frac{16 a^5 b^3}{45}, \\
&\frac{128}{45} a^3 b^3 \left(\frac{a^2}{5} + \frac{b^2}{7}\right) a_1 + \frac{128}{45 \times 35} a^5 b^5 (a^2 + b^2) a_2 \\
&\qquad + \frac{128}{45 \times 7} \left(\frac{11}{5} a^2 + \frac{1}{3} b^2\right) a_3 = \frac{16 a^3 b^5}{45}.
\end{aligned}\right\} \tag{4}$$

Solving these, we obtain

$$\left.\begin{aligned} a_1 &= \frac{35(9a^4+130a^2b^2+9b^4)}{16(45a^6+509a^4b^2+509a^2b^4+45b^6)};\\ a_2 &= \frac{105(9a^2+b^2)}{16(45a^6+509a^4b^2+509a^2b^4+45b^6)};\\ a_3 &= \frac{105(a^2+9b^2)}{16(45a^6+509a^4b^2+509a^2b^4+45b^6)}. \end{aligned}\right\} \tag{5}$$

If we retain only one coefficient a_1, by putting

$$u \approx u_1 = a_1(x^2-a^2)(y^2-b^2), \tag{6}$$

we then readily find

$$a_1 = \frac{5}{8(a^2+b^2)}. \tag{7}$$

In order to judge the relative degree of precision of the solutions constructed here and in § 12 we proceed in the following fashion. As is known the torque M is defined by the formula

$$M = 2\int_{-a}^{a}\int_{-b}^{b} \psi \,dx\,dy = 4G\theta\int_{-a}^{a}\int_{-b}^{b} u\,dx\,dy,$$

which can be represented in the form:

$$M = (2a)^3(2b)G\theta k_1\left(\frac{b}{a}\right). \tag{8}$$

(See S. P. Timoshenko [1], § 78 whence come all the following data on the exact solution of our problem.)

The maximum tangential stress in the rectangular section occurs, as is known, half-way along the side. Let $b>a$. The greatest tangential stress τ can be represented in the form

$$\tau = G\theta 2ak\left(\frac{b}{a}\right). \tag{9}$$

We now construct new approximate solutions u_1, u_3, u_6 by retaining one, three or six terms respectively in series (8) § 12:

$$u_1(x, y)=\frac{16a^2b^2}{\pi^4(a^2+b^2)}\sin\frac{\pi x}{a}\sin\frac{\pi y}{b},$$

$$u_3(x, y)=\frac{16a^2b^2}{\pi^4}\left\{\frac{1}{a^2+b^2}\sin\frac{\pi x}{a}\sin\frac{\pi y}{b}\right.$$
$$+\frac{1}{3(9a^2+b^2)}\sin\frac{\pi x}{a}\sin\frac{3\pi y}{b}$$
$$\left.+\frac{1}{3(a^2+9b^2)}\sin\frac{3\pi x}{a}\sin\frac{\pi y}{b}\right\},$$

$$u_6(x, y)=\frac{16a^2b^2}{\pi^4}\left\{\frac{1}{a^2+b^2}\sin\frac{\pi x}{a}\sin\frac{\pi y}{b}\right. \tag{10}$$
$$+\frac{1}{3(9a^2+b^2)}\sin\frac{\pi x}{a}\sin\frac{3\pi y}{b}+\frac{1}{3(a^2+9b^2)}\sin\frac{3\pi x}{a}\sin\frac{\pi y}{b}$$
$$+\frac{1}{5(25a^2+b^2)}\sin\frac{\pi x}{a}\sin\frac{5\pi y}{b}+\frac{1}{81(a^2+b^2)}\sin\frac{3\pi x}{a}\sin\frac{3\pi y}{b}$$
$$\left.+\frac{1}{5(a^2+25b^2)}\sin\frac{5\pi x}{a}\sin\frac{\pi y}{b}\right.$$

With the aid of these approximate solutions we shall evaluate two mechanical characteristics of the torsional problem —the torque and the maximal tangential stress—and we shall compare the approximate solutions obtained with the known exact solutions. Denoting the approximate values of M by M_1, M_3, M_6, we have

$$\left.\begin{aligned} M_1&=\frac{256G\theta a^3b^3}{\pi^6(a^2+b^2)},\\ M_3&=\frac{256G\theta a^3b^3}{\pi^6}\left[\frac{1}{(a^2+b^2)}+\frac{1}{9(9a^2+b^2)}+\frac{1}{9(a^2+9b^2)}\right.\\ M_6&=\frac{256G\theta a^3b^3}{\pi^6}\left[\frac{730}{729(a^2+b^2)}+\frac{1}{9(9a^2+b^2)}+\frac{1}{9(a^2+9b^2)}\right.\\ &\quad\left.\left.+\frac{1}{25(25a^2+b^2)}+\frac{1}{25(a^2+25b^2)}\right]\right. \end{aligned}\right\} \tag{11}$$

We recall that in formulae (10) and (11) the lengths of the sides of the rectangle are denoted by a and b and

not by $2a$ and $2b$; however in the following this is unimportant.

We denote by $k_{1,i}(\gamma)$, $k^{(i)}(\gamma)$, $i=1, 3, 6$ the approximate values of $k_1(\gamma)$ and $k(\gamma)$, $\gamma=\dfrac{b}{a}$ evaluated according to the approximate solution u_i [formula (10)]. We readily find

$$k_{1,1}(\gamma)=\frac{256}{\pi^6}\frac{\gamma^2}{1+\gamma^2},$$

$$k_{1,3}(\gamma)=\frac{256\gamma^2}{\pi^6}\left[\frac{1}{1+\gamma^2}+\frac{1}{9(9+\gamma^2)}+\frac{1}{9(1+9\gamma^2)}\right],$$

$$k_{1,6}(\gamma)=\frac{256\gamma^2}{\pi^6}\left[\frac{1}{1+\gamma^2}+\frac{1}{9(9+\gamma^2)}+\frac{1}{9(1+9\gamma^2)}+\frac{1}{25(25+\gamma^2)}+\frac{1}{25(1+25\gamma^2)}\right].$$

and

$$k^{(1)}(\gamma)=\frac{32}{\pi^3}\frac{\gamma^2}{1+\gamma^2},$$

$$k^{(3)}(\gamma)=\frac{32\gamma^2}{\pi^3}\left[\frac{1}{1+\gamma^2}-\frac{1}{3(\gamma^2+9)}+\frac{1}{1+9\gamma^2}\right],$$

$$k^{(6)}(\gamma)=\frac{32\gamma^2}{\pi^3}\left[\frac{1}{1+\gamma^2}-\frac{1}{3(\gamma^2+9)}+\frac{1}{1+9\gamma^2}+\frac{1}{5(\gamma^2+25)}+\frac{1}{1+25\gamma^2}\right].$$

Tables 1 and 2 below allow us to compare the exact values of the quantities k_1 and k with their approximate values.

TABLE 1

γ	k_1	$k_{1,1}$	$k_{1,3}$	$k_{1,6}$
1	0·1406	0·133	0·139	0·1401
2	0·229	0·213	0·225	0·228
3	0·263	0·240	0·258	0·261
4	0·281	0·251	0·273	0·278
5	0·291	0·256	0·281	0·287
∞	0·333	0·266	0·299	0·311

TABLE 2

γ	k	$k^{(1)}$	$k^{(3)}$	$k^{(6)}$
1	0·675	0·516	0·585	0·613
2	0·930	0·826	0·831	0·870
3	0·985	0·929	0·870	0·931
4	0·997	0·971	0·865	0·954
5	0·999	0·992	0·854	0·961
∞	1·000	1·032	0·803	1·012

Tables 1 and 2 show that the approximate solutions give fairly good values of the torque but considerably poorer values of the maximal tangential stress (this particularly applies to u_3 (x, y)). This is occasioned by the fact that the precision of the evaluation of the torque M depends upon the rapidity *of the convergence of series* (8) § 12 *in the mean*, and the precision of the evaluation of τ, upon the rapidity *of the uniform convergence of the series of derivatives*. It is readily seen that series (8) § 12 converges in the mean fairly rapidly but that the uniform convergence of the derivatives of this series is extremely slow.

For comparison we give the values $k_{1,1}$ (γ), $k_{1,3}$ (γ), $k^{(1)}$ (γ) and $k^{(3)}$ (γ) evaluated from the approximate solutions (3) and (6). We retain the notation of tables 1 and 2.

TABLE 3

γ	k_1	$k_{1,1}$	$k_{1,3}$
1	0·1406	0·139	0·1404
2	0·229	0·222	0·228
3	0·263	0·250	0·263
4	0·281	0·261	0·279
5	0·291	0·267	0·290
∞	0·333	0·278	0·311

TABLE 4

γ	k_1	$k^{(1)}$	$k^{(3)}$
1	0·675	0·625	0·703
2	0·930	1·000	0·951
3	0·985	1·125	0·982
4	0·997	1·176	0·969
5	0·999	1·202	0·951
∞	1·000	1·250	0·875

Comparison with tables 1 and 2 shows that for an approximation of the same order the algebraic polynomials give better approximations than the trigonometric for the torque,

and worse approximations for the maximum tangential stress. However we cannot assert that this circumstance will persist for approximations with a greater number of terms.

2. Solution by the method of orthogonal projections. According to the method of orthogonal projections we put

$$-\nabla u = \mathbf{v};$$

$\mathbf{v}$ is a two-dimensional vector with the components $v_x = -\dfrac{\partial u}{\partial x}$ and $v_k = -\dfrac{\partial u}{\partial y}$. Equation (1) gives

$$\nabla \cdot \mathbf{v} = 1. \tag{12}$$

We construct some vector $\mathbf{V}$ which satisfies equation (12). We can put $\mathbf{V} = (x, 0)$, i. e.

$$V_x = x, \quad V_y = 0.$$

We now select a system of vectors $\mathbf{w}_n$ which satisfy the equation

$$\nabla \cdot \mathbf{w}_n = 0. \tag{13}$$

A complete system of such vectors can be obtained by choosing the $\mathbf{w}_n$ such that their components have the form of products of powers of x and y. However we do not require all such vectors.

The fact is that $u(x,y)$ is an even function both with respect to x and with respect to y, therefore the component $v_x = -\dfrac{\partial u}{\partial x}$ must be odd with respect to x and even with respect to y, and the component $v_y = -\dfrac{\partial u}{\partial y}$ is odd with respect to y and even with respect to x. Obviously the component of the vector $\mathbf{w} = \mathbf{V} - \mathbf{v}$ must also possess analogous properties. Since we only require the vectors $\mathbf{w}_n$ in order to approximate to the vector $\mathbf{w}$, we can subject them to the same conditions and we can put

$$\left.\begin{aligned} &\mathbf{w}_1 = (x; -y),\ \mathbf{w}_2 = (x^3;\ -3x^2y),\ \mathbf{w}_3 = (3xy^2;\ -y^3), \\ &\mathbf{w}_4 = (x^5; -5x^4y),\ \mathbf{w}_5 = (x^3y^2; -x^2y^3),\ \mathbf{w}_6 = (5x^4y; -y^5), \ldots \end{aligned}\right\} \tag{14}$$

We restrict ourselves to the first three terms of this sequence and we put

$$\mathbf{w} \approx \mathbf{w}^{(3)} = a_1\mathbf{w}_1 + a_2\mathbf{w}_2 + a_3\mathbf{w}_3$$

or, *in extenso*

$$w_x \approx w_x^{(3)} = a_1 x + a_2 x^3 + 3a_3 xy^2,$$

$$w_y \approx w_y^{(3)} = -a_1 y - 3a_2 x^2 y - a_3 y^3.$$

The coefficients, a_1, a_2, a_3 must be found from the equation

$$\|\mathbf{V} - \mathbf{w}\|^2 = \min,$$

or

$$\int_{-a}^{a}\int_{-b}^{b}[(x - a_1 x - a_2 x^3 - 3a_3 xy^2)^2 + (a_1 y + 3a_2 x^2 y + a_3 y^3)^2]\,dx\,dy$$

is to be a minimum.

By performing the integration and then differentiating the left-hand side with respect to a_1, a_2, a_3 and equating the derivatives to zero we obtain the system of equations:

$$\begin{aligned}
&\left(\frac{4a^3b}{3} + \frac{4ab^3}{3}\right)a_1 + \left(\frac{4a^5b}{5} + \frac{4a^3b^3}{3}\right)a_2 \\
&\quad + \left(\frac{4a^3b^3}{3} + \frac{4ab^5}{5}\right)a_3 = \frac{4a^3b}{3}; \quad \left(\frac{4a^5b}{5} + \frac{4a^3b^3}{3}\right)a_1 \\
&+ \left(\frac{4a^7b}{7} + \frac{12a^5b^3}{5}\right)a_2 + \left(\frac{4a^5b^3}{5} + \frac{4a^3b^5}{5}\right)a_3 = \frac{4a^5b}{5}; \qquad (15) \\
&\left(\frac{4a^3b^3}{3} + \frac{4ab^5}{5}\right)a_1 + \left(\frac{4a^5b^3}{5} + \frac{4a^3b^5}{5}\right)a_2 \\
&\qquad + \left(\frac{12a^3b^5}{5} + \frac{4ab^7}{7}\right)a_3 = \frac{4a^3b^3}{3}.
\end{aligned}$$

Solving this system, we obtain

$$\left.\begin{aligned}
a_1 &= \frac{a^2}{a^2+b^2} + \frac{21a^2b^2(a^4-b^4),}{7a^8+114a^6b^2+214a^4b^4+114a^2b^6+7b^8} \\
a_2 &= -\frac{7b^2(35a^4+38a^2b^2+3b^4),}{6(7a^8+114a^6b^2+214a^4b^4+114a^2b^6+7b^8)}, \\
a_3 &= \frac{7a^2(3a^4+38a^2b^2+35b^4).}{6(7a^8+114a^6b^2+214a^4b^4+114a^2b^6+7b^8)}.
\end{aligned}\right\} \qquad (16)$$

Having obtained a solution constructed by the method of orthogonal projections, it is possible to estimate the error of the solutions constructed both by the Ritz method and by the method of orthogonal projections. In order to simplify the exposition we restrict ourselves to the case of a square ($a=b$). By virtue of inequality (14) § 50 it is possible to put $\delta=-\|\mathrm{V}-\mathrm{w}^{(3)}\|^2$ in the general formula (4) § 48. We have

$$\|\mathrm{V}-\mathrm{w}^{(3)}\|^2=\left\|\mathrm{V}-\sum_{k=1}^{3}a_k\mathrm{w}_k\right\|^2,$$

where this time the coefficients a_k are defined by formulae (16). Further

$$\left\|\mathrm{V}-\sum_{k=1}^{3}a_k\mathrm{w}_k\right\|^2=\left(\mathrm{V}-\sum_{j=1}^{3}a_j\mathrm{w}_j,\ \mathrm{V}-\sum_{k=1}^{3}a_k\mathrm{w}_k\right)$$

$$=\|\mathrm{V}\|^2-2\sum_{k=1}^{3}a_k(\mathrm{V},\ \mathrm{w}_k)+\sum_{j,\,k=1}^{3}a_ja_k(\mathrm{w}_j,\ \mathrm{w}_k). \qquad (17)$$

The coefficients a_k satisfy the system

$$\sum_{j=1}^{3}a_j(\mathrm{w}_j,\ \mathrm{w}_k)=(\mathrm{V},\ \mathrm{w}_k);\quad k=1,\ 2,\ 3.$$

Multiplying by a_k and summing, we obtain,

$$\sum_{j,\,k=1}^{3}a_ja_k(\mathrm{w}_j,\ \mathrm{w}_k)=\sum_{k=1}^{3}a_k(\mathrm{V},\ \mathrm{w}_k).$$

Substituting this in (1), we find

$$\left\|\mathrm{V}-\sum_{k=1}^{3}a_k\mathrm{w}_k\right\|^2=\|\mathrm{V}\|^2-\sum_{k=1}^{3}a_k(\mathrm{V},\ \mathrm{w}_k).$$

But

$$\|\mathrm{V}\|^2=\int_{-a}^{a}\int_{-a}^{a}x^2\mathrm{d}x\mathrm{d}y=\frac{4}{3}a^4,$$

the coefficients a_k are given by formulae (16) and the quantities $(\mathrm{V},\ \mathrm{w}_k)$ are the right-hand sides of equations (15). Using these data we readily find

$$\left\|\mathrm{V}-\sum_{k=1}^{3}a_k\mathrm{V}_k\right\|^2=\frac{76}{135}a^4=0{\cdot}5630a^4.$$

We now evaluate $F(u_3)$, where u_3 is the approximate solution (3) constructed by the Ritz method. By formula (8) § 48 we have

$$F(u_3) = -[a_1(f, \varphi_1) + a_2(f, \varphi_2) + a_3(f, \varphi_3)];$$

The scalar products here are the terms not involving the a_j in system (4) and the coefficients a_1, a_2, a_3 are defined by formula (5). Putting $a = b$, we obtain

$$F(u_3) = -\frac{1400}{2493} a^4 = -0{\cdot}5616a^4.$$

Denoting the approximate solution constructed by the Ritz method or by the method of orthogonal projections by u_3, we have

$$\|u - u_3\| \leq a^2\gamma(0{\cdot}5630 - 0{\cdot}5616) = a^2\gamma(0{\cdot}0014) = 0{\cdot}037a^2,$$

which gives a relative error in the energy of the order of 6%.

Using the approximate solutions given by the method of orthogonal projections it is possible to estimate the quantity $\tau_{\max}$; the corresponding values of the quantity $k(\gamma)$ are given in table 5.

TABLE 5

γ	1	2	3	4	5	∞
Exact Solution	0·675	0·930	0·985	0·997	0·999	1·000
Solution by the method of orthogonal projections	0·694	0·982	1·047	1055	1·050	1·000

Comparison with tables 2 and 4 shows that the accuracy of the evaluation of $\tau_{\max}$ by the method of orthogonal projections is somewhat higher than by the Ritz method and is close to the accuracy given by the series in the eigenfunctions of the Laplacian operator (§ 12).

3. Application of Trefftz's method. In equation (1) we put $u = P - \frac{x^2}{2}$ then

$$\nabla^2 p = 0, \quad p|_S = \frac{x^2}{2},$$

where S is the boundary of the rectangle. In applying the Ritz method we restrict ourselves to one term so that the approximate solution will have the form

$$p_1(x, y) = a_1\varphi_1(x, y),$$

where φ_1 is a function which is harmonic in the rectangle. We take

$$\varphi_1 = \cos\frac{\pi x}{2a}\cosh\frac{\pi y}{2a}.$$

The system of equations of Trefftz's method (§ 55) reduces to the one equation

$$a_1\int_S \varphi_1\frac{\partial\varphi_1}{\partial n}\,dS = \int_S \frac{x^2}{2}\frac{\partial\varphi_1}{\partial n}\,dS.$$

Carrying out the integrations, we find

$$a_1 = -\frac{16a^3}{\pi^3}\frac{1}{\cosh\dfrac{\pi b}{2a}}$$

the approximate value of the function u, to within a constant term, equals

$$u \approx -\frac{x^2}{2} - \frac{16a^2}{\pi^3}\frac{\cosh\dfrac{\pi y}{2a}}{\cosh\dfrac{\pi b}{2a}}\cos\frac{\pi x}{2a}.$$

We give a table of the values of $k(\gamma)$ corresponding to this approximate solution (see the second column of table 6). For the calculations we have used the formula

$$\tau = |\tau_{yz}|_{x=a,\,y=0} = 2G\theta\left|\frac{\partial u}{\partial x}\right|_{x=a,\,y=0}$$

As we see the result is considerably more precise than that from the Ritz method or from the method of orthogonal projections. This is connected with the selection of the coordinate function φ_1, which approximates well to the required harmonic function $p(x, y)$.

Rather worse results are obtained if we use harmonic polynomials, i.e. if we put

$$u=\tilde{u}-\frac{1}{4}(x^2+y^2),\quad \nabla^2\tilde{u}=0,\quad \tilde{u}|_S=\frac{1}{4}(x^2+y^2).$$

In using Trefftz's method to find the function $\tilde{u}$ we introduce harmonic polynomials as coordinate functions; since the function $\tilde{u}$ is even with respect to each of the coordinates x and y we can restrict ourselves to even polynomials only. We put

$$\tilde{u}=a_1(x^2-y^2)+a_2(x^4-6x^2y^2+y^4)$$
$$+a_3(x^6-15x^4y^2+15x^2y^4-y^6). \tag{18}$$

Without dwelling on the details of the calculations we give the values of $\tau_{\max}$ obtained with the aid of the approximate solution (18) (see the third column of Table 6).

TABLE 6

γ	Solution		
	exact	Trefftz's (16)	Trefftz's (18)
1	0·675	0·676	0·694
2	0·930	0·930	0·931
3	0·985	0·985	0·971
4	0·997	0·997	0·973
5	0·999	0·999	0·970
∞	1·000	1·000	1·000

§ 66. Bending of a rectangular plate, rigidly fixed at the rim

The problem consists of integrating the biharmonic equation

$$\nabla^4 w=\frac{q}{D}=p, \tag{1}$$

where q is the load intensity and D is the rigidity of the plate, under the boundary conditions

$$w|_S = 0, \quad \left.\frac{\partial w}{\partial n}\right|_S = 0. \tag{2}$$

We denote the lengths of the sides of the plate by $2a$ and $2b$. We take the coordinate axes parallel to the sides of the plate and we place the origin of coordinates at its centre.

As was proved in § 27, the problem reduces to the problem of finding the minimum of the functional

$$\int_{-a}^{a} dx \int_{-b}^{b} \{(\nabla^2 w)^2 - 2pw\} dy \tag{3}$$

for the set of functions which satisfy boundary conditions (2). This last problem can be solved by the Ritz method. In order to simplify the calculations we shall suppose that the load is distributed uniformly, so that $p = \text{const}$.

As the coordinate functions we can take polynomials of the form

$$(x^2 - a^2)^2 (y^2 - b^2)^2 (a_1 + a_2 x^2 + a_3 y^2 + \ldots); \tag{4}$$

we omit odd powers of x and y because $w(x, y)$ is obviously symmetric about the coordinate axes.

We restrict ourselves to three terms in expression (4): using the notation

$$\varphi_1 = (x^2 - a^2)^2 (y^2 - b^2)^2, \quad \varphi_2 = x^2 \varphi_1, \quad \varphi_3 = y^2 \varphi_1,$$

we shall have

$$w \approx w_3 = a_1 \varphi_1 + a_2 \varphi_2 + a_3 \varphi_3.$$

The equations of the Ritz method in this case have the form

$$\sum_{k=1}^{3} (\nabla^2 \varphi_k, \nabla^2 \varphi_m) a_k = p(1, \varphi_m), \quad m = 1, 2, 3. \tag{5}$$

By performing the necessary calculations, we transform the system (5) to the form

$$\left.\begin{aligned}
&\left(\gamma^2+\frac{1}{\gamma^2}+\frac{4}{7}\right)a_1+\left(\frac{1}{7}+\frac{1}{11\gamma^4}\right)b^2a_2+\left(\frac{1}{7}+\frac{\gamma^4}{11}\right)a^2a_3\\
&\qquad=\frac{7}{128a^2b^2}\,p\\
&\left(\frac{\gamma^2}{7}+\frac{1}{11\gamma^2}\right)a_1+\left(\frac{3}{7}+\frac{4}{143\gamma^4}+\frac{4}{77\gamma^2}\right)b^2a_2+\frac{1}{77}(1+\gamma^4)a^2a_3\\
&\qquad=\frac{1}{128a^2b^2}\,p\\
&\left(\frac{1}{7\gamma^2}+\frac{\gamma^2}{11}\right)a_1+\frac{1}{77}\left(1+\frac{1}{\gamma^4}\right)b^2a_2+\left(\frac{3}{7}+\frac{3\gamma^4}{143}+\frac{4\gamma^2}{77}\right)a^2a_3\\
&\qquad=\frac{1}{128a^2b^2}\,p
\end{aligned}\right\}\qquad(6)$$

Here $\gamma=\dfrac{b}{a}$. The equations of system (6) have been obtained from (5) by dividing the left-and right-hand sides by various common factors.

If the plate is square then $\gamma=1$, and equations (6) assume the simpler form

$$\left.\begin{aligned}
&\frac{18}{7}a_1+\frac{18}{77}a^2a_2+\frac{18}{77}a^2a_3=\frac{7}{128a^4}p;\\
&\frac{18}{77}a_1+\frac{502}{1001}a^2a_2+\frac{2}{77}a^2a_3=\frac{1}{128a^4}p;\\
&\frac{18}{77}a_1+\frac{2}{77}a^2a_2+\frac{502}{1001}a^2a_3=\frac{1}{128a^4}p.
\end{aligned}\right\}\qquad(7)$$

Solving this system, we obtain

$$w\approx w_3=\frac{p}{a^4}(x^2-a^2)^2(y^2-a^2)^2\left(0{\cdot}02067+0{\cdot}0038\frac{x^2+y^2}{a^2}\right).\qquad(8)$$

We obtain for the deflexion at the centre of the plate

$$w\Big|_{\substack{x=0\\y=0}}=0{\cdot}02067pa^4.\qquad(9)$$

If in the approximate expression for w we restrict ourselves to one term and put

$$w\approx w_1=a_1\varphi_1=a_1(x^2-a^2)^2(y^2-b^2)^2,$$

then, instead of system (6), we obtain the single equation

$$\left(\gamma^2+\frac{1}{\gamma^2}+\frac{4}{7}\right)a_1=\frac{7}{128a^2b^2}p,$$

whence

$$a_1=\frac{49}{128(7a^4+7b^4+4a^2b^2)}p, \tag{10}$$

which gives the value

$$w\Big|_{\substack{x=0\\y=0}}=\frac{49a^4b^4}{128(7a^4+7b^4+4a^2b^2)}. \tag{11}$$

for the deflexion at the centre of the plate.

In the case of a square ($a=b$) we obtain

$$w\Big|_{\substack{x=0\\y=0}}=0{\cdot}02127pa^4. \tag{12}$$

Formulae (10) and (12) are given in L. S. Leibenzon's monograph [1]. We now apply the method of the anharmonic residue to our problem. (The calculations in this section which are connected with the application of the method of anharmonic residue are taken in part from a paper by Z. Kh. Rafal'son [2].)

For simplicity we suppose that the plate is square, with side a. For the function $q(x, y)$ which satisfies the equation $\nabla^2 q=p$ we take

$$q=\frac{p}{4}(x^2+y^2).$$

We introduce a set of harmonic polynomials which, as can be proved, is complete in the subspace of functions which are harmonic within the square and satisfy the condition

$$\int_{-a}^{a}\int_{-a}^{a}u^2\mathrm{d}x\mathrm{d}y<\infty. \tag{13}$$

However, we do not need all of this set. In fact from symmetry, it is obvious that the required function $w(x, y)$ is even both

in x and in y; moreover it is unchanged by interchanging the arguments x and y. Therefore we need only retain those harmonic polynomials which satisfy the same symmetry properties.

We now write down the first three such polynomials:

$$N_1=1, \quad N_2=x^4-6x^2y^2+y^4,$$
$$N_3=x^8-28x^6y^2+70x^4y^4-28x^2y^6+y^8.$$

By orthonormalizing them in the square domain we obtain three new harmonic polynomials:

$$M_1=\frac{1}{2a}, \qquad M_2=\frac{1}{2}\sqrt{\frac{7}{13}}\left[\frac{1}{2a}+\frac{15}{8a^5}(x^4-6x^2y^2+y^4)\right],$$
$$M_3=0\cdot356359\left(M_1-\sqrt{\left(\frac{7}{13}\right)}\times4\cdot76724M_2-1\cdot40625a^{-9}N_3\right).$$

The corresponding Fourier coefficients of the function $q(x, y)$ are

$$q_1=(q, M_1)=\frac{pa^3}{3}, \quad q_2=(q, M_2)=-0\cdot1397pa^3,$$

$$q_3=(q, M_3)=0\cdot0014pa^3.$$

By virtue of theorem 1 § 56 we can take for δ (formula [4] § 48), the value

$$\delta=-\left(\|q\|^2-\sum_{i=1}^{3}q_i^2\right)=-0\cdot02491p^2a^6.$$

We now evaluate the quantity $F(w_3)$, where F is defined by formula (3), and w_3 by formula (8). By formula (8) § 48

$$F(w_3)=-\sum_{i=1}^{3}a_i(f, \varphi_i), \qquad f=p.$$

From formula (7) we have

$$a_1=0\cdot02067\frac{p}{a^4}, \quad a_2=a_3=0\cdot0038\frac{p}{a^6}.$$

Further, as is easily seen,

$$(f, \varphi_1) = (p, \varphi_1) = p \int_{-a}^{a} \int_{-a}^{a} (x^2 - a^2)^2 (y^2 - a^2)^2 \mathrm{d}x\mathrm{d}y = \frac{256}{225} p a^{10},$$

$$(f, \varphi_2) = (f, \varphi_3) = p \int_{-a}^{a} \int_{-a}^{a} x^2 (x^2 - a^2)^2 (y^2 - a^2)^2 \mathrm{d}x\mathrm{d}y \frac{256}{7 \times 225} p a^{12}.$$

Hence $F(w_3) = -0{\cdot}02486\ p^2 a^6$ and

$$|w - w_3| \leq p a^3 \sqrt{(0{\cdot}00005)} = 0{\cdot}0071 p a^3. \tag{14}$$

Now it is also easy to estimate the absolute value of the difference $w - w_3$. Let $u(x, y)$ be any function which is twice continuously differentiable in the closed square and vanishes together with its first derivative on the sides of the square. Suppose, for example, that the point (x, y) lies in the first quadrant. We have

$$u(x, y) = \int_{x}^{a} \int_{y}^{a} \frac{\partial^2 u(\xi, \eta)}{\partial \xi \partial \eta} \mathrm{d}\xi \mathrm{d}\eta.$$

From Cauchy's inequality

$$u^2(x, y) \leq \int_{x}^{a} \int_{y}^{a} 1^2 \mathrm{d}\xi \mathrm{d}\eta \int_{x}^{a} \int_{y}^{a} \left(\frac{\partial^2 u}{\partial \xi \partial \eta} \right)^2 \mathrm{d}\xi \mathrm{d}\eta$$

$$= (a - x)(a - y) \int_{x}^{a} \int_{y}^{a} \left(\frac{\partial^2 u}{\partial \xi \partial \eta} \right)^2 \mathrm{d}\xi \mathrm{d}\eta.$$

In the first quadrant the maximum value of each of the differences $a - x$ and $a - y$ equals a. Hence we readily obtain the inequality

$$u^2(x, y) \leq a^2 \int_{-a}^{a} \int_{-a}^{a} \left(\frac{\partial^2 u}{\partial x \partial y} \right)^2 \mathrm{d}x\mathrm{d}y$$

which is valid, as is easily seen, in the entire square. Recalling that in our problem

$$\|u\|^2=\int_{-a}^{a}\int_{-a}^{a}\left[\left(\frac{\partial^2 u}{\partial x^2}\right)^2+2\left(\frac{\partial^2 u}{\partial x\partial y}\right)^2+\left(\frac{\partial^2 u}{\partial y^2}\right)^2\right]dxdy,$$

we find $u^2\leq\frac{a^2}{2}\|u\|^2$; applying this inequality to the function $w-w_3$ and using the inequality (14) we finally obtain

$$|w-w_3|\leq\frac{0{\cdot}0071}{\sqrt{2}}pa^4=0{\cdot}005pa^4. \tag{15}$$

§ 67. Bending of semicircular plate elastically supported at the rim

For a rigidly fixed rim, the conditions

$$w|_L=0,\quad \left.\frac{\partial w}{\partial n}\right|_L=0,$$

are satisfied, and at a freely supported rim, the conditions

$$w|_L=0,\quad \left[\nabla^2 w-\frac{1-\sigma}{\varrho}\frac{\partial w}{\partial n}\right]_L=0.$$

We shall say that the rim of the plate is *elastically supported* if at this boundary the conditions

$$w|_L=0, \tag{1}$$

$$\left[\nabla^2 w-\frac{1-\sigma}{\varrho}\frac{\partial w}{\partial n}+k\frac{\partial w}{\partial n}\right]_L=0, \tag{2}$$

are satisfied, where k is a positive constant. Over the disc we assume as usual, that the Sophie Germain–Lagrange equation

$$\nabla^4 w=\frac{q}{D} \tag{3}$$

is satisfied.

The biharmonic operator is symmetric and positive-bounded-below for the set of functions which satisfy conditions (1) and (2). In order to prove this we construct the scalar product

$$(\nabla^4 w_1, w_2) = \int_S \int w_2 \nabla^4 w_1 \mathrm{d}S,$$

in which both the functions w_1 and w_2 are subject to conditions (1) and (2).

By Green's formula (formula 10 § 2)

$$\int_S \int w_2 \nabla^4 w_1 \mathrm{d}S = \int_S \int \nabla^2 w_1 \nabla^2 w_2 \mathrm{d}S + \int_L \left(w_2 \frac{\partial \nabla^2 w_1}{\partial n} - \nabla^2 w_1 \frac{\partial w_2}{\partial n} \right) \mathrm{d}s,$$

or, if we use boundary conditions (1) and (2),

$$(\nabla^4 w_1, w_2) = \int_S \int \nabla^2 w_1 \nabla^2 w_2 \mathrm{d}S - \int_L \left(\frac{1-\sigma}{\varrho} - k \right) \frac{\partial w_1}{\partial n} \frac{\partial w_2}{\partial n} \mathrm{d}s. \quad (4)$$

The right-hand side of equation (4) is unchanged by interchanging the functions w_1 and w_2 therefore $(\nabla^4 w_1, w_2) = (w_1, \nabla^4 w_2)$ and our biharmonic operator is symmetric.

Now we put $w_1 = w_2 = w$ in (4). Then

$$(\nabla^4 w, w) = \int_S \int (\nabla^2 w)^2 \mathrm{d}S - \int_L \frac{1-\sigma}{\varrho} \left(\frac{\partial w}{\partial n} \right)^2 \mathrm{d}s + k \int_L \left(\frac{\partial w}{\partial n} \right)^2 \mathrm{d}s$$

$$\geq \int_S \int (\nabla^2 w)^2 \mathrm{d}S - \int_L \frac{1-\sigma}{\varrho} \left(\frac{\partial w}{\partial n} \right)^2 \mathrm{d}s. \quad (5)$$

Formulae (20) and (21) § 27 are valid for any functions which vanish on the boundary L. Comparison of these formulae gives

$$\int_S \int (\nabla^2 w)^2 \mathrm{d}S - \int_L \frac{1-\sigma}{\varrho} \left(\frac{\partial w}{\partial n} \right)^2 \mathrm{d}s$$

$$= \int_S \int \left\{ \left(\frac{\partial^2 w}{\partial x^2} \right)^2 + 2\sigma \frac{\partial^2 w}{\partial x^2} \frac{\partial^2 w}{\partial y^2} + \left(\frac{\partial^2 w}{\partial y^2} \right)^2 + 2(1-\sigma) \left(\frac{\partial^2 w}{\partial x \partial y} \right)^2 \right\} \mathrm{d}S. \quad (6)$$

Now by inequality (5)

$$(\nabla^4 w,\ w) \geq \int_S \int \left\{ \left(\frac{\partial^2 w}{\partial x^2} \right)^2 + 2\sigma \frac{\partial^2 w}{\partial x^2} \frac{\partial^2 w}{\partial y^2} + \left(\frac{\partial^2 w}{\partial y^2} \right)^2 + 2(1-\sigma) \left(\frac{\partial^2 w}{\partial x \partial y} \right)^2 \right\} \mathrm{d}s. \qquad (7)$$

The remainder of the argument proceeds as in subsection 4 § 27, and leads to inequality (26) § 27. This also implies that our operator is positive-bounded-below.

From what has been said it follows that the problem of a disc which is elastically supported at the rim can be posed as the problem of finding the minimum of the functional

$$\int_S \int \left\{ (\nabla^2 w)^2 + 2(1-\sigma) \left[\left(\frac{\partial^2 w}{\partial x \partial y} \right)^2 - 2 \frac{\partial^2 w}{\partial x^2} \frac{\partial^2 w}{\partial y^2} \right] \right\} \mathrm{d}S + k \int_L \left(\frac{\partial w}{\partial n} \right)^2 \mathrm{d}s - 2 \int_S \int \frac{q}{D} w \, \mathrm{d}S;$$

under boundary condition (1). Condition (2), as is readily seen, is natural.

As an illustration we consider a semicircular plate elastically supported at the rim. We assume that the radius of the plate equals unity; the plate is shown in Fig. 18. In order to simplify the calculations we assume that $\sigma = 0$, $k = 1$, $\frac{q}{D} = 1$. We shall use the Ritz method.

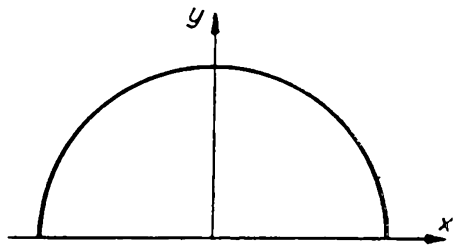

FIG. 18

As the coordinate functions, we select functions of the form

$$x^{2k}y^l(1-x^2-y^2);$$
$$k=0,\ 1,\ 2,\ \ldots;\quad l=1,\ 2,\ 3,$$

which satisfy the principal boundary condition (1); we do not include odd powers of x since the deflexion of the plate is symmetric about the y-axis. Restricting ourselves to terms for which $2k+l\leq 4$ we obtain

$$w_6=\sum_{k=1}^{6}a_k\varphi_k,$$

where

$$\varphi_1=y(1-x^2-y^2),\ \varphi_2=y\varphi_1,\ \varphi_3=x^2\varphi_1,\ \varphi_4=y^2\varphi_1,$$
$$\varphi_5=x^2y\varphi_1,\quad \varphi_6=y^3\varphi_1.$$

The coordinate functions do not satisfy the natural condition (2) and it is necessary to take the system of equations of the Ritz method in the form of (8_2) § 14. Since $\sigma=0$, in our case,

$$[\varphi_i,\ \varphi_k]=\int_S\int\left[\nabla^2\varphi_i\nabla^2\varphi_k+2\frac{\partial^2\varphi_i}{\partial x\partial y}\frac{\partial^2\varphi_k}{\partial x\partial y}-\frac{\partial^2\varphi_i}{\partial x^2}\frac{\partial^2\varphi_k}{\partial y^2}\right.$$
$$\left.-\frac{\partial^2\varphi_k}{\partial x^2}\frac{\partial^2\varphi_i}{\partial y^2}\right]\mathrm{d}S+k\int_L\frac{\partial\varphi_i}{\partial n}\frac{\partial\varphi_k}{\partial n}\mathrm{d}s;$$

Performing the necessary calculations, we obtain the system

$$26{\cdot}1994a_1+21{\cdot}3333a_2+6{\cdot}43557a_3+18{\cdot}8496a_4+4{\cdot}2667a_5$$
$$+17{\cdot}0667a_6=0{\cdot}266667,$$
$$21{\cdot}3333a_1+24{\cdot}0855a_2+4{\cdot}87619a_3+23{\cdot}1619a_4+4{\cdot}02517a_5$$
$$+21{\cdot}9911a_6=0{\cdot}130900,$$
$$6{\cdot}43557a_1+4{\cdot}87619a_2+5{\cdot}54858a_3+3{\cdot}92699a_4+3{\cdot}25079a_5$$
$$+3{\cdot}25079a_6=0{\cdot}038095,$$
$$18{\cdot}8496a_1+23{\cdot}1619a_2+3{\cdot}92699a_3+24{\cdot}3473a_4+3{\cdot}05714a_5$$
$$+24{\cdot}3810a_6=0{\cdot}076190,$$
$$4{\cdot}26667a_1+4{\cdot}02517a_2+3{\cdot}25079a_3+3{\cdot}05714a_4+2{\cdot}36928a_5$$
$$+3{\cdot}18080a_6=0{\cdot}016362,$$
$$17{\cdot}0667a_1+21{\cdot}9911a_2+3{\cdot}25079a_3+24{\cdot}3810a_4+3{\cdot}18080a_5$$
$$+25{\cdot}4076a_6=0{\cdot}049087.$$

Solving this system, we find the following values of the coefficients:

$$a_1 = 0{\cdot}02217; \quad a_2 = -0{\cdot}01145; \quad a_3 = -0{\cdot}01317$$

$$a_4 = 0{\cdot}000887; \quad a_5 = 0{\cdot}00759; \quad a_6 = -0{\cdot}01083.$$

§ 68. Evaluation of the eigenvalues of an ordinary second-order differential equation

With the purpose of illustrating the computational process, we consider the equation

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\sqrt{(1+x)}\frac{\mathrm{d}u}{\mathrm{d}x}\right)+\lambda u=0 \tag{1}$$

under the boundary conditions

$$u(0)=u(1)=0. \tag{2}$$

In seeking a solution by the Ritz method we take the coordinate functions

$$\varphi_k(x)=(1-x)x^k, \qquad k=1,\ 2,\ \ldots.$$

We restrict ourselves to the first three functions

$$\varphi_1(x)=(1-x)x,\ \varphi_2(x)=(1-x)x^2,\ \varphi_3(x)=(1-x)x^3.$$

Applying the method of § 32 (we omit the details of the calculation) we find that λ is defined by the equation

$$\begin{vmatrix} 0{\cdot}404757774-\frac{1}{30}\lambda; & 0{\cdot}216156130-\frac{1}{60}\lambda; & 0{\cdot}135002283-\frac{1}{105}\lambda \\ 0{\cdot}216156130-\frac{1}{60}\lambda; & 0{\cdot}510789821-\frac{1}{105}\lambda; & 0{\cdot}675363337-\frac{1}{168}\lambda \\ 0{\cdot}135002282-\frac{1}{105}\lambda; & 0{\cdot}675363337-\frac{1}{168}\lambda; & 1{\cdot}023822057-\frac{1}{252}\lambda \end{vmatrix} = 0 \tag{3}$$

We obtain the crudest and most simple approximation to the first eigenvalue by equating the first-order diagonal minor to zero:

$$0{\cdot}404757774-\frac{1}{30}\lambda=0, \tag{4}$$

This corresponds to using only one coordinate function $\varphi_1(x)$ in the Ritz method. From equation (4) we obtain

$$\lambda_1^{(1)}=12{\cdot}14273.$$

(The superscript denotes the degree of the approximation and the subscript, the number of the eigenvalue.)

We obtain the second approximation by equating the second-order minor to zero

$$\begin{vmatrix} 0{\cdot}404757774-\frac{1}{30}\lambda; & 0{\cdot}216156130-\frac{1}{60}\lambda \\ 0{\cdot}216156130-\frac{1}{60}\lambda; & 0{\cdot}510789821-\frac{1}{105}\lambda \end{vmatrix}=0; \tag{5}$$

The smallest root of equation (5) equals

$$\lambda_1^{(2)}=12{\cdot}12781,$$

which gives a more precise approximation to the least eigenvalue of problem (1)–(2). Solving equation (3) by Newton's method we obtain the still more precise value

$$\lambda_1^{(3)}=12{\cdot}12255. \tag{6}$$

Each of the numbers $\lambda_1^{(1)}$, $\lambda_1^{(2)}$, $\lambda_1^{(3)}$ is greater than the eigenvalue λ_1. We obtain a lower bound to λ_1 by using the method mentioned in subsection 2 § 62 which is based upon the reduction of the given problem to an integral equation.

Equation (1) under boundary conditions (2) is equivalent to the integral equation

$$u(x)-\lambda\int_0^1 G(x,\ s)u(s)\mathrm{d}s=0, \tag{7}$$

where $G(x, s)$ is Green's function for the given problem. (For more detail concerning the construction of the Green function cf. e. g. V. I. Smirnov [2], Ch. IV, § 1.) In order to construct this function we integrate the equation

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\sqrt{(1+x)}\frac{\mathrm{d}v}{\mathrm{d}x}\right)=0;$$

its general solution is $v=C_1\sqrt{(1+x)}+C_2$. We find two solutions of which the first vanishes at $x=0$, and the second at $x=1$:

$$v_1=A(\sqrt{(1+x)}-1), \qquad v_2=B(\sqrt{(1+x)}-\sqrt{(2)}).$$

We shall find Green's function if we consider the constants A and B as functions of the parameter s and subject them to the requirements that at $x=s$

$$v_1=v_2, \quad \frac{\partial v_1}{\partial x}-\frac{\partial v_2}{\partial x}=-\frac{1}{\sqrt{(1+s)}}.$$

This gives the equations

$$A(\sqrt{(1+s)}-1)=B(\sqrt{(1+s)}-\sqrt{(2)});$$
$$\frac{A}{2\sqrt{(1+s)}}-\frac{B}{2\sqrt{(1+s)}}=-\frac{1}{\sqrt{(1+s)}}.$$

Hence, for example

$$B=\frac{2(\sqrt{(1+s)}-1)}{\sqrt{(2)}-1}$$

Now

$$G(x, s)=\frac{2(\sqrt{(1+s)}-1)(\sqrt{(1+x)}-\sqrt{(2)})}{\sqrt{(2)}-1}, \quad x\geq s.$$

Since Green's function is symmetric

$$G(x, s)=\frac{2(\sqrt{(1+x)}-1)(\sqrt{(1+s)}-\sqrt{(2)})}{\sqrt{(2)}-1}, \quad x\leq s.$$

Taking $G(x,s)$ as the kernel of integral equation (7) we construct by the usual rules the second integrated kernel; we then obtain, for $x\leq s$

$$G_2(x, s)=\frac{4}{3(3-2\sqrt{2})}\left\{(\sqrt{(2)}-\sqrt{(1+s)})\left[(\sqrt{(2-1)})\times\right.\right.$$

$$\times\left(-2-x-\frac{x^2}{2}-2\gamma(1+x)+x\gamma(1+x)\right)+4(\gamma(2)-\gamma(1+x))\Big]$$
$$+(\gamma(1+x)-1)\Big[(\gamma(2)-1)\left(2+s+\frac{s^2}{2}\right)$$
$$+2\gamma(2)\gamma(1+s)-\gamma(2s)\gamma(1+s)-\frac{11}{2}(\gamma(1+s)-1)\Big]\Big\};$$

the value of $G_2(x,\ s)$ for $x \geq s$ is obtained, by symmetry, by interchanging the arguments x and s. From formula (*) § 62 we find

$$a_2=\int_0^1\int_0^1 G^2(x,\ s)\mathrm{d}x\mathrm{d}s,$$

or, since the kernel is symmetric,

$$a_2=2\int_0^1 \mathrm{d}s\int_0^s G^2(x,\ s)\mathrm{d}x,$$

In our case this gives

$$a_2=0{\cdot}0076414951.$$

Similarly

$$a_4=2\int_0^1 \mathrm{d}s\int_0^s G_2^2(x,\ s)\mathrm{d}x=0{\cdot}50092905\times10^{-4}.$$

By formula (4) § 62 we obtain a lower approximate value for the least eigenvalue

$$\lambda_{11}=\frac{1}{\gamma(a_2)}=11{\cdot}4395997;\quad \lambda_{12}=\frac{1}{\sqrt[4]{\gamma(a_4)}}=11{\cdot}886553.$$

Taking a more precise approximation we finally obtain

$$11{\cdot}88655\leq\lambda_1\leq12{\cdot}12255.$$

By taking the arithmetic mean of the lower and upper approximate values we obtain the approximate value

$$\lambda_1=12{\cdot}00455$$

which has relative error of less than 1%.

§ 69. Eigenvibrations of a beam of variable cross-section

The equation for the free vibrations of a beam of variable cross-section has the form

$$E\frac{\partial^2}{\partial x^2}\left[I(x)\frac{\partial^2 z}{\partial x^2}\right]+\varrho S(x)\frac{\partial^2 z}{\partial t^2}=0. \tag{1}$$

Here, as usual, the x-axis is directed along the axis of the beam, $z(x, t)$ is its transverse displacement, $I(x)$ and $S(x)$ are the geometrical moment of inertia and the area of cross-section at the abscissa x, E and ϱ are Young's modulus and the density of the beam material. To be explicit we suppose that one end of the beam is fixed and that the other is free. We denote the length of the rod by L and place the coordinate origin at its fixed end. Then the boundary conditions of our problem can be written thus:

$$z|_{x=0}=0, \quad \left.\frac{\partial z}{\partial x}\right|_{x=0}=0; \tag{2}$$

$$\left.\frac{\partial^2 z}{\partial x^2}\right|_{x=L}=0, \quad \left.\frac{\partial^3 z}{\partial x^3}\right|_{x=L}=0. \tag{3}$$

As usual we shall seek a solution in the form

$$z(x, t)=u(x)\sin(\sqrt{\lambda}t+\alpha), \ \alpha=\text{const}. \tag{4}$$

Then equation (1) and boundary conditions (2) and (3) transform to the following

$$E\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left[I(x)\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}\right]-\varrho\lambda S(x)u=0; \tag{5}$$

$$u(0)=0, \quad u'(0)=0; \tag{6}$$

$$u''(L)=0, \quad u'''(L)=0. \tag{7}$$

As proved in § 35, the least eigenvalue λ_1 of our problem equals the minimum of the integral

$$\int_0^L Eu\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left[I(x)\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}\right]\mathrm{d}x=E\int_0^L I(x)\left[\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}\right]^2\mathrm{d}x \tag{8}$$

for the set of functions satisfying boundary conditions (6) and (7) and the supplementary condition

$$\int_0^L \varrho S(x)u^2(x)\mathrm{d}x = 1. \tag{9}$$

In finding the minimum of integral (8) by the Ritz method, we take as the coordinate functions the eigenfunctions of the operator $\frac{d^4u}{dx^4}$ under boundary conditions (6) and (7). These functions form a complete orthonormal system. As is known (cf. V. N. Faddeyeva [1], [2]) the functions mentioned have the form

$$\left.\begin{aligned} \varphi_{2m-1}(\xi) &= \frac{\sin \alpha_k \xi}{\sin \frac{\alpha_k}{2}} + \frac{\cosh \alpha_k \xi}{\cosh \frac{\alpha_k}{2}}, \quad k = 2m-1, \\ \varphi_{2m}(\xi) &= \frac{\cos \alpha_k \xi}{\cos \frac{\alpha_k}{2}} - \frac{\sinh \alpha_k \xi}{\sinh \frac{\alpha_k}{2}}, \quad k = 2m, \end{aligned}\right\} \tag{10}$$

where the α_k are the roots of the equation

$$\cos \alpha \cosh \alpha = 1;$$

(the values of α_k are given, for example, in a paper by V. N. Faddeyeva [1], table 7). In formula (10) the interval is $-\frac{1}{2} \leq \xi \leq \frac{1}{2}$ and conditions (6) refer to the end $\xi = -\frac{1}{2}$ and conditions (7), to the end $\xi = +\frac{1}{2}$. In order to use the functions (10) in our problem, obviously, we must put

$$\xi = \frac{2x-L}{2L}. \tag{11}$$

Putting approximately

$$u(x) \approx \sum_{k=1}^{n} a_k \psi_k(x), \quad \psi_k(x) = \varphi_k\left(\frac{2x-L}{2L}\right),$$

we obtain, corresponding to what was said in § 35, the following equation for determining λ:

$$\begin{vmatrix} A_{11}-\lambda B_{11}; & A_{12}-\lambda B_{12}; & \ldots . & A_{1n}-\lambda B_{1n} \\ A_{21}-\lambda B_{21}; & A_{22}-\lambda B_{22}; & \ldots . & A_{2n}-\lambda B_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ A_{n1}-\lambda B_{n1}; & A_{n2}-\lambda B_{n2}; & \ldots . & A_{nn}-\lambda B_{nn} \end{vmatrix} = 0, \qquad (12)$$

where

$$A_{ik}=E\int_0^L I(x)\psi_i''(x)\psi_k''(x)\mathrm{d}x, \qquad B_{ik}=\int_0^L \varrho S(x)\psi_i(x)\psi_k(x)\mathrm{d}x. \qquad (13)$$

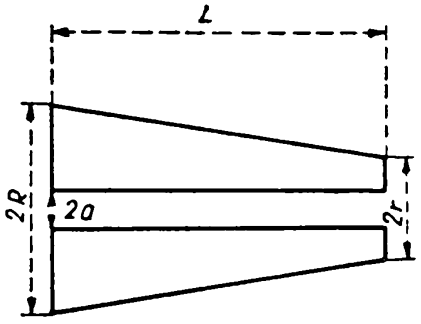

FIG. 19

As an example we consider a uniform tube of uniform material in the form of a hollow truncated cone; an axial section of this tube is shown in Fig. 19. In this case

$$I(x)=\frac{\pi}{4}(y^4-a^4),$$

$$S(x)=\pi(y^2-a^2),$$

where

$$y=R-\frac{R-r}{L}x=R-\left(\xi+\frac{1}{2}\right)(R-r).$$

If we make the substitution $\xi=\dfrac{2x-L}{2L}$ in (13) the coefficients A_{ik} and B_{ik} assume the form:

$$A_{ik}=\frac{\pi E}{4L^3}\left\{\left[\frac{1}{16}(R+r)^4-a^4\right]I_{ik}^{(0)}-\frac{1}{2}(R+r)(R-r)I_{ik}^{(1)}+\right.$$

$$+\frac{3}{2}(R+r)^2(R-r)^2I_{ik}^{(2)}-2(R+r)(R-r)^3I_{ik}^{(3)}+(R-r)^4I_{ik}^{(4)}\Big\}; \quad (14)$$

$$B_{ik}=\pi\varrho L\left\{\left[\frac{(R+r)^2}{4}-a^2\right]\tilde{I}_{ik}^{(0)}-\frac{1}{8}(R^2-r^2)\tilde{I}_{ik}^{(1)}+(R-r)^2\tilde{I}_{ik}^{(2)}\right\}. \quad (15)$$

Here $I_{ik}^{(1)}$, ..., $\tilde{I}_{ik}^{(2)}$ are integrals defined by the formulae

$$\left.\begin{aligned} I_{ik}^{(n)}&=\int_{-\frac{1}{2}}^{\frac{1}{2}}\xi^n\varphi_i''(\xi)\varphi_k''(\xi)\mathrm{d}\xi, \quad n=0,\ 1,\ 2,\ 3,\ 4;\\ \tilde{I}_{ik}^{(n)}&=\int_{-\frac{1}{2}}^{\frac{1}{2}}\xi^n\varphi_i(\xi)\varphi_k(\xi)\mathrm{d}\xi, \quad n=0,\ 1,\ 2. \end{aligned}\right\} \quad (16)$$

We obtain the following group of formulae for the $I_{ik}^{(n)}$, i and k having various values.

I. $i=k-$odd

$$I_{kk}^{(0)}=\alpha_k^4; \quad I_{kk}^{(1)}=-2\alpha_k^2\cot^2\frac{\alpha_k}{2};$$

$$I_{kk}^{(2)}=\frac{\alpha_k^4}{12}+\alpha_k\cot\frac{\alpha_k}{2}-\frac{\alpha_k^2}{2}\cot^2\frac{\alpha_k}{2};$$

$$I_{kk}^{(3)}=-\frac{3}{2}\left(\alpha_k^2\cot^2\frac{\alpha_k}{2}-4\alpha_k\cot\frac{\alpha_k}{2}+4\right);$$

$$I_{kk}^{(4)}=\frac{\alpha_k^4}{80}+\frac{1}{4}\left(6\alpha_k\cot\frac{\alpha_k}{2}-\alpha_k^2\cot^2\frac{\alpha_k}{2}-6\right).$$

II. $i=k$ – even

$$I_{kk}^{(0)}=\alpha_k^4; \quad I_{kk}^{(1)}=-2\alpha_k^2\tan^2\frac{\alpha_k}{2};$$

$$I_{kk}^{(2)}=\frac{\alpha_k^4}{12}-\alpha_k\tan\frac{\alpha_k}{2}-\frac{\alpha_k^2}{2}\tan^2\frac{\alpha_k}{2};$$

$$I_{kk}^{(3)}=-\frac{3}{2}\left(\alpha_k^2\tan^2\frac{\alpha_k}{2}+4\alpha_k\tan\frac{\alpha_k}{2}+4\right);$$

$$I_{kk}^{(4)}=\frac{\alpha_k^4}{80}-\frac{1}{4}\left(6\alpha_k\tan\frac{\alpha_k}{2}+\alpha_k^2\tan^2\frac{\alpha_k}{2}+6\right).$$

III. $i \neq k$, i – odd, k – odd.

$$I_{ik}^{(0)} = 0; \quad I_{ik}^{(1)} = -\frac{8\alpha_i^3\alpha_k^3}{(\alpha_i^2+\alpha_k^2)^2}\cot\frac{\alpha_i}{2}\cot\frac{\alpha_k}{2};$$

$$I_{ik}^{(2)} = \frac{8\alpha_i^2\alpha_k^2}{(\alpha_i^2-\alpha_k^2)^2}\left\{\alpha_i\alpha_k\cot\frac{\alpha_i}{2}\cot\frac{\alpha_k}{2}\right.$$

$$\left. -\frac{(3\alpha_i^2+\alpha_k^2)}{\alpha_i^2-\alpha_k^2}\alpha_k\cot\frac{\alpha_k}{2}+\frac{(\alpha_i^2+3\alpha_k^2)}{\alpha_i^2-\alpha_k^2}\alpha_i\cot\frac{\alpha_i}{2}\right\};$$

$$I_{ik}^{(3)} = \frac{6\alpha_i^2\alpha_k^2}{(\alpha_i^2+\alpha_k^2)^2}\left\{\frac{4(\alpha_i^4-6\alpha_i^2\alpha_k^2+\alpha_k^4)}{(\alpha_i^2+\alpha_k^2)^2}\right.$$

$$-\frac{2}{\alpha_i^2+\alpha_k^2}\left[(\alpha_k^2-3\alpha_i^2)\alpha_k\cot\frac{\alpha_k}{2}+(\alpha_i^2-3\alpha_k^2)\alpha_i\cot\frac{\alpha_i}{2}\right]$$

$$\left.-\alpha_i\alpha_k\cot\frac{\alpha_i}{2}\cot\frac{\alpha_k}{2}\right\};$$

$$I_{ik}^{(4)} = \frac{4\alpha_i^2\alpha_k^2}{(\alpha_i^2-\alpha_k^2)^2}\left\{\alpha_i\alpha_k\cot\frac{\alpha_i}{2}\cot\frac{\alpha_k}{2}-\frac{12(\alpha_i^4+6\alpha_i^2\alpha_k^2+\alpha_k^4)}{(\alpha_i^2-\alpha_k^2)^2}\right.$$

$$\left.+\frac{3}{\alpha_i^2-\alpha_k^2}\left[(\alpha_i^2+3\alpha_k^2)\alpha_i\cot\frac{\alpha_i}{2}-(\alpha_k^2+3\alpha_i^2)\alpha_k\cot\frac{\alpha_k}{2}\right]\right\};$$

IV. $i \neq k$, i even, k even.

$$I_{ik}^{(0)} = 0; \quad I_{ik}^{(1)} = -\frac{8\alpha_i^3\alpha_k^3}{(\alpha_i^2+\alpha_k^2)^2}\tan\frac{\alpha_i}{2}\tan\frac{\alpha_k}{2};$$

$$I_{ik}^{(2)} = \frac{8\alpha_i^2\alpha_k^2}{(\alpha_i^2-\alpha_k^2)^2}\left\{\alpha_i\alpha_k\tan\frac{\alpha_i}{2}\tan\frac{\alpha_k}{2}\right.$$

$$\left.+\frac{\alpha_k^2+3\alpha_i^2}{\alpha_i^2-\alpha_k^2}\alpha_k\tan\frac{\alpha_k}{2}-\frac{\alpha_i^2+3\alpha_k^2}{\alpha_i^2-\alpha_k^2}\alpha_i\tan\frac{\alpha_i}{2}\right\};$$

$$I_{ik}^{(3)} = \frac{6\alpha_i^2\alpha_k^2}{(\alpha_i^2+\alpha_k^2)^2}\left\{\frac{4(\alpha_i^4-6\alpha_i^2\alpha_k^2+\alpha_k^4)}{(\alpha_i^2+\alpha_k^2)^2}\right.$$

$$-\frac{2}{\alpha_i^2+\alpha_k^2}\left[(\alpha_k^2-3\alpha_i^2)\alpha_k\tan\frac{\alpha_i}{2}+(\alpha_i^2-3\alpha_k^2)\alpha_i\tan\frac{\alpha_k}{2}\right]$$

$$\left.-\alpha_i\alpha_k\tan\frac{\alpha_i}{2}\tan\frac{\alpha_k}{2}\right\};$$

$$I_{ik}^{(4)}=\frac{4\alpha_i^2\alpha_k^2}{(\alpha_i^2-\alpha_k^2)^2}\left\{\alpha_i\alpha_k\tan\frac{\alpha_i}{2}\tan\frac{\alpha_k}{2}\right.$$

$$-\frac{3}{\alpha_i^2-\alpha_k^2}\left[(\alpha_k^2+3\alpha_i^2)\alpha_k\tan\frac{\alpha_k}{2}-(\alpha_i^2+3\alpha_k^2)\alpha_i\tan\frac{\alpha_i}{2}\right]$$

$$\left.-\frac{12(\alpha_i^4+6\alpha_i^2\alpha_k^2+\alpha_k^4)}{(\alpha_i^2-\alpha_k^2)^2}\right\}.$$

V. i odd, k even.

$$I_{ik}^{(0)}=0;\quad I_{ik}^{(1)}=-\frac{8\alpha_i^3\alpha_k^3}{(\alpha_i^2-\alpha_k^2)^2}\tan\frac{\alpha_k}{2}\cot\frac{\alpha_i}{2};$$

$$I_{ik}^{(2)}=\frac{8\alpha_i^2\alpha_k^2}{(\alpha_i^2+\alpha_k^2)^2}\left\{\alpha_i\alpha_k\cot\frac{\alpha_i}{2}\tan\frac{\alpha_k}{2}\right.$$

$$\left.+\frac{(\alpha_k^2-3\alpha_i^2)\alpha_k\tan\frac{\alpha_k}{2}-(\alpha_i^2-3\alpha_k^2)\alpha_i\cot\frac{\alpha_i}{2}}{\alpha_i^2+\alpha_k^2}\right\};$$

$$I_{ik}^{(3)}=-\frac{6\alpha_i^2\alpha_k^2}{(\alpha_i^2-\alpha_k^2)^2}\left\{\alpha_i\alpha_k\cot\frac{\alpha_i}{2}\tan\frac{\alpha_k}{2}\right.$$

$$-\frac{2}{\alpha_i^2-\alpha_k^2}\left[(\alpha_i^2+3\alpha_k^2)\alpha_i\cot\frac{\alpha_i}{2}-(\alpha_k^2+3\alpha_i^2)\alpha_k\tan\frac{\alpha_k}{2}\right]$$

$$\left.+\frac{4(\alpha_i^4+6\alpha_i^2\alpha_k^2+\alpha_k^4)}{(\alpha_i^2-\alpha_k^2)^2}\right\};$$

$$I_{ik}^{(4)}=\frac{4\alpha_i^2\alpha_k^2}{(\alpha_i^2+\alpha_k^2)^2}\left\{\alpha_i\alpha_k\cot\frac{\alpha_i}{2}\tan\frac{\alpha_k}{2}+\frac{3}{\alpha_i^2+\alpha_k^2}\left[(\alpha_k^2-3\alpha_i^2)\alpha_k\tan\frac{\alpha_k}{2}\right.\right.$$

$$\left.\left.-(\alpha_i^2-3\alpha_k^2)\alpha_i\cot\frac{\alpha_i}{2}\right]+\frac{12(\alpha_i^4-6\alpha_i^2\alpha_k^2+\alpha_k^4)}{(\alpha_i^2+\alpha_k^2)^2}\right\};$$

The $\tilde{I}_{ik}^{(n)}$ have the values

$$\tilde{I}_{kk}^{(0)}=1;\quad \tilde{I}_{ik}^{(0)}=0,\quad i\neq k.$$

$$\tilde{I}_{ik}^{(1)}=-\frac{I_{ik}^{(1)}}{\alpha_i^2\alpha_k^2};\quad \tilde{I}_{ik}^{(2)}=\frac{I_{ik}^{(2)}}{\alpha_i^2\alpha_k^2}.$$

For the numerical calculations we adopt the following dimensions for the tube; $L=4000$ mm, $R=400$ mm, $r=200$ mm, $a=100$ mm.

By putting $n=1, 2, 3$ in the approximate expression

$$u \approx \sum_{k=1}^{n} a_k u_k,$$

we obtain the following equations for determining λ:

1) $0{\cdot}8526-0{\cdot}1917\times10^{3}\lambda\frac{\varrho}{E}=0.$

2) $$\begin{vmatrix} 0{\cdot}8526-0{\cdot}1917\times10^{3}\lambda\frac{\varrho}{E}; & -1{\cdot}5341+0{\cdot}0632\times10^{3}\lambda\frac{\varrho}{E} \\ -1{\cdot}5341+0{\cdot}0632\times10^{3}\lambda\frac{\varrho}{E}; & 39{\cdot}4262-0{\cdot}3526\times10^{3}\varrho\frac{\varrho}{E} \end{vmatrix}=0,$$

3) $$\begin{vmatrix} 0{\cdot}8526-0{\cdot}1917\times10^{3}\varrho\frac{\varrho}{E}; & -1{\cdot}5341+0{\cdot}0632\times19^{3}\lambda\frac{\varrho}{E}; & 1{\cdot}1077+0{\cdot}0037\times10^{3}\lambda\frac{\varrho}{E} \\ -1{\cdot}5341+0{\cdot}0632\times10^{3}\lambda\frac{\varrho}{E}; & 39{\cdot}4262-0{\cdot}3526\times10^{3}\lambda\frac{\varrho}{E}; & -26{\cdot}614+0{\cdot}0865\times10^{3}\lambda\frac{\varrho}{E} \\ 1{\cdot}1077+0{\cdot}0037\times10^{3}\lambda\frac{\varrho}{E}; & -26{\cdot}614+0{\cdot}0865\times10^{3}\lambda\frac{\varrho}{E}; & 156{\cdot}59-0{\cdot}3168\times10^{3}\lambda\frac{\varrho}{E} \end{vmatrix}=0,$$

the least roots of which are

$$\lambda_1^{(1)}=0{\cdot}4446\times10^{-8}\frac{E}{\sigma};\quad \lambda_1^{(2)}=0{\cdot}4226\times10^{-8}\frac{E}{\varrho};$$

$$\lambda_1^{(3)}=0{\cdot}4171\times10^{-8}\frac{E}{\varrho}.$$

As is clear the successive values of λ_1 converge fairly rapidly: the relative error on passing from $\lambda_1^{(1)}$ to $\lambda_1^{(3)}$ is about 5% and on passing from $\lambda_1^{(2)}$ to $\lambda_1^{(3)}$, about 1·3%. For the second eigenvalue equations (2) and (3) give the approximate values

$$\lambda_2^{(2)}=11{\cdot}628\times10^{-8}\frac{E}{\varrho},\quad \lambda_2^{(3)}=11{\cdot}074\times10^{-8}\frac{E}{\varrho};$$

and the relative error on passing from $\lambda_2^{(2)}$ to $\lambda_2^{(3)}$ approximately equals 5%.

We shall now try to find lower bounds for the eigenfrequencies of the vibrating beam. In order to simplify the calculations we shall discuss the case of a solid tube ($a=0$).

From the Ritz method, the eigenvalues are again determined from a system of the form (12). All that is necessary is that the coefficients A_{ik} and B_{ik} be replaced by the new coefficients $\overline{A}_{ik}$ and $\overline{B}_{ik}$; their evaluation requires no special labour. As we can readily satisfy ourselves,

$$\overline{A}_{ik}=A_{ik}+\frac{\pi E a^4}{4L^3}I_{ik}^{(0)};\quad \overline{B}_{ik}=B_{ik}+\pi\varrho L a^2 I_{ik}^{(0)}.$$

Again putting $n=3$ we obtain for the least value of λ the upper bound

$$\lambda_1\approx 0{\cdot}34498\times 10^{-8}\frac{E}{\varrho}. \tag{17}$$

In order to estimate a lower bound for λ_1, we reduce our problem, by the method of § 62, to an integral equation with a symmetric kernel. This kernel has the form

$$K(x,t)=\sqrt{\{S(x)S(t)\}}G(x,t),$$

where $G(x,t)$ is Green's function for our problem and $S(x)$ is the cross-sectional area of the beam at abscissa x. Green's function $G(x,t)$ is defined as that solution of the differential equation

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(I(x)\frac{\mathrm{d}^2G}{\mathrm{d}x^2}\right)=0,$$

which is continuous together with its first and second-order derivatives, when the third derivative undergoes a jump at $x=t$ defined by

$$\left.\frac{\partial^3 G}{\partial x^3}\right|_{t=x-0}-\left.\frac{\partial^3 G}{\partial x^3}\right|_{t=x+0}=\frac{1}{I(x)}.$$

Performing the necessary calculations, we find that for $x\leq t$

$$K(x,t)=\frac{2(R-Ct)}{3R^2(R-Cx)}[3Rtx^2-(R+2Ct)x^3];\quad C=\frac{R-r}{L}.$$

It is now possible to evaluate the second trace of the kernel:

$$A_2=2\int_0^L \mathrm{d}t\int_0^t K^2(x,t)\,\mathrm{d}x=8{\cdot}979548\times 10^{16}.$$

The parameter of the integral equation is $\lambda\frac{\varrho}{E}$. By formula (4) § 62 we obtain a lower bound for λ_1:

$$\lambda_1 \approx 0{\cdot}33371 \times 10^{-8}\frac{E}{\varrho}. \tag{18}$$

The values (17) and (18) have been obtained for the dimensions given for figure 19.

The arithmetic mean of values (17) and (18) gives

$$\lambda_1 = 0{\cdot}3390 \times 10^{-8}\frac{E}{\varrho}$$

with a relative error of less than 1·5%.

§ 70. Radial eigenvibrations of an elastic cylinder

The lowest eigenfrequency ω_1 of the vibrations of an elastic body whose surface is load-free is defined by the relation

$$\gamma\omega_1^2 = \min\frac{2E(\mathbf{u})}{H(\mathbf{u})}, \tag{1}$$

where γ is the density of the medium, $E(\mathbf{u})$ is the potential energy of deformation:

$$2E(\mathbf{u}) = \int_\Omega \{\lambda\theta^2 + 2\mu(\varepsilon_x^2 + \varepsilon_y^2 + \varepsilon_z^2) + \mu(\gamma_{xy}^2 + \gamma_{xz}^2 + \gamma_{yz}^2)\}\mathrm{d}\Omega, \tag{2}$$

$$\theta = \varepsilon_x + \varepsilon_y + \varepsilon_z$$

and

$$H(\mathbf{u}) = \int_\Omega (u_x^2 + u_y^2 + u_z^2)\mathrm{d}\Omega = \|\mathbf{u}\|^2. \tag{3}$$

The condition for the absence of loads on the boundary of the body is natural. Thus there is no necessity to subject $\mathbf{u}$ to any boundary conditions. However $\mathbf{u}$ must be subject to the relations

$$\int_\Omega \mathbf{u}\mathrm{d}\Omega = 0, \quad \int_\Omega \mathbf{r}\times\mathbf{u}\mathrm{d}\Omega = 0, \tag{4}$$

expressing the absence of rigid displacements.

We shall consider the case where the elastic medium occupies a circular cylinder of radius R and height h, and we suppose that the vibrations are radial, so that $\mathbf{u}$ in cylindrical coordinates ϱ, θ, z, has only two components u_ϱ and u_z independent of the angle θ. In formulae (2) and (3) we replace the cartesian coordinates by cylindrical coordinates. We then obtain

$$2E(\mathbf{u})=\int_\Omega\{\lambda(\varepsilon_\varrho+\varepsilon_\vartheta+\varepsilon_z)^2+2\mu(\varepsilon_\varrho^2+\varepsilon_\vartheta^2+\varepsilon_z^2)$$
$$+\mu(\gamma_{\varrho\vartheta}^2+\gamma_{\varrho z}^2+\gamma_{\vartheta z}^2)\}\mathrm{d}\Omega, \tag{5}$$
$$H(\mathbf{u})=\int_\Omega(u_\varrho^2+u_\vartheta^2+u_z^2)\mathrm{d}\Omega, \tag{6}$$

where, under the conditions of radial symmetry

$$u_\vartheta=0,\quad \frac{\partial u_\varrho}{\partial\vartheta}=\frac{\partial u_z}{\partial\vartheta}=0,$$

$$\varepsilon_\varrho=\frac{\partial u_\varrho}{\partial\varrho},\quad \varepsilon_\vartheta=\frac{u_\varrho}{\varrho},\quad \varepsilon_z=\frac{\partial u_z}{\partial z},\quad \gamma_{\varrho z}=\frac{\partial u_\varrho}{\partial z}+\frac{\partial u_z}{\partial\varrho},\quad \gamma_{\varrho\vartheta}=\gamma_{\vartheta z}=0.$$

Substituting these in (5) and (6) and carrying out the integration with respect to θ, we find

$$2E(\mathbf{u})=4\pi E_1(\mathbf{u});\quad H(\mathbf{u})=2\pi H_1(\mathbf{u}),$$

where

$$2E_1(\mathbf{u})=\int_0^R\varrho\mathrm{d}\varrho\int_0^h\left\{\lambda\left(\frac{\partial u_\varrho}{\partial\varrho}+\frac{u_\varrho}{\varrho}+\frac{\partial u_z}{\partial z}\right)^2\right.$$
$$\left.+2\mu\left[\left(\frac{\partial u_\varrho}{\partial\varrho}\right)^2+\frac{u_\varrho^2}{\varrho^2}+\left(\frac{\partial u_z}{\partial z}\right)^2\right]+\mu\left(\frac{\partial u_\varrho}{\partial z}+\frac{\partial u_z}{\partial\varrho}\right)^2\right\}\mathrm{d}z, \tag{7}$$
$$H_1(\mathbf{u})=\int_0^R\varrho\mathrm{d}\varrho\int_0^h(u_\varrho^2+u_z^2)\mathrm{d}z. \tag{8}$$

Thus,

$$\gamma\omega_1^2=\min\frac{2E_1(\mathbf{u})}{H_1(\mathbf{u})}. \tag{9}$$

In finding the minimum of ratio (9) by the Ritz method, we take approximate displacement components u_ϱ and u_z in the form

$$u_\varrho=\varrho(a_0+a_1z+a_2\varrho^2+a_3z^2), \tag{10}$$
$$u_z=b_0+b_1z+b_2\varrho^2+b_3z^2.$$

The factor ϱ is introduced into u_ϱ because, as is clear from the imposed symmetry, $u_\varrho = 0$ at $\varrho = 0$.

The coefficients in (10) are not arbitrary, but are connected by the relations (4) which give

$$a_0 = -3\left(\frac{a_1 h}{6} + \frac{a_2 R^2}{5} + \frac{a_3 h^2}{9}\right),$$

$$b_0 = -\left(\frac{b_1 h}{2} + \frac{b_2 R^2}{2} + \frac{b_3 h^2}{2}\right), \tag{11}$$

$$a_1 = \frac{6}{h^2}\left(\frac{b_2 R^2}{5} - \frac{a_3 h^3}{6}\right).$$

Relations (11) allow us to eliminate the coefficients a_0, b_0, a_1; this gives us

$$u_\varrho = a_2\left(\varrho^3 - \frac{3R^2}{5}\varrho\right) + a_3\left(\frac{h^2}{6}\varrho - hz\varrho + z^2\varrho\right) + b_2\left(\frac{6R^2}{5h^2} z\varrho - \frac{3R^2}{5h}\varrho\right), \tag{12}$$

$$u_z = b_1\left(z - \frac{h}{2}\right) + b_2\left(\varrho^2 - \frac{R^2}{2}\right) + b_3\left(z^2 - \frac{h^2}{3}\right).$$

We introduce the displacement vectors

$$\boldsymbol{\varphi}_1 = \left(0;\ z - \frac{h}{2}\right),\quad \boldsymbol{\varphi}_2 = \left(\varrho^3 - \frac{3R^2}{5}\varrho;\ 0\right),$$

$$\boldsymbol{\varphi}_3 = \left(\frac{6R^2}{5h^2}\varrho z - \frac{3R^2}{5h}\varrho;\ \varrho^2 - \frac{R^2}{2}\right), \tag{13}$$

$$\boldsymbol{\varphi}_4 = \left(\frac{h^2}{6}\varrho - h\varrho z + z^2\varrho;\ 0\right),\quad \boldsymbol{\varphi}_5 = \left(0;\ z^2 - \frac{h^2}{3}\right);$$

In formulae (13), the component along the ϱ-axis is written first and that along the z-axis is written second. With the aid of the vectors $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_5$ it is possible to write formulae (12) in the more convenient form

$$\mathbf{u} = b_1\boldsymbol{\varphi}_1 + b_2\boldsymbol{\varphi}_2 + b_3\boldsymbol{\varphi}_3 + b_4\boldsymbol{\varphi}_4 + b_5\boldsymbol{\varphi}_5. \tag{14}$$

We introduce the notation

$$2E_1(\mathbf{u}, \mathbf{v}) = \int_0^R \varrho \, d\varrho \int_0^h \left\{ \lambda \left(\frac{\partial u_\varrho}{\partial \varrho} + \frac{u_\varrho}{\varrho} + \frac{\partial u_z}{\partial z} \right) \left(\frac{\partial v_\varrho}{\partial \varrho} + \frac{v_\varrho}{\varrho} + \frac{\partial v_z}{\partial z} \right) \right.$$

$$\left. + 2\mu \left(\frac{\partial u_\varrho}{\partial \varrho} \frac{\partial v_\varrho}{\partial \varrho} + \frac{u_\varrho v_\varrho}{\varrho^2} + \frac{\partial u_z}{\partial z} \frac{\partial v_z}{\partial z} \right) + \mu \left(\frac{\partial u_\varrho}{\partial z} + \frac{\partial u_z}{\partial \varrho} \right) \left(\frac{\partial v_\varrho}{\partial z} + \frac{\partial v_z}{\partial \varrho} \right) \right\} dz;$$

$$H_1(\mathbf{u}, \mathbf{v}) = \int_0^R \varrho \, d\varrho \int_0^k (u_\varrho v_\varrho + u_z v_z) dz.$$

As follows from the general results of § 33, the equation for determining the quantity $\gamma\omega_1^2 = \varkappa$ has the form:

$$\begin{vmatrix} 2E_1(\varphi_1, \varphi_1) - \varkappa H_1(\varphi_1, \varphi_1); & 2E_1(\varphi_1, \varphi_2) - \varkappa H_1(\varphi_1, \varphi_2); \ldots & 2E_1(\varphi_1, \varphi_n) - \varkappa H_1(\varphi_1, \varphi_n) \\ 2E_1(\varphi_1, \varphi_2) - \varkappa H_1(\varphi_1, \varphi_2); & 2E_1(\varphi_2, \varphi_2) - \varkappa H_1(\varphi_2, \varphi_2); \ldots & 2E_1(\varphi_2, \varphi_n) - \varkappa H_1(\varphi_2, \varphi_n) \\ \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots \\ 2E_1(\varphi_1, \varphi_n) - \varkappa H_1(\varphi_1, \varphi_n); & 2E_1(\varphi_2, \varphi_n) - \varkappa H_1(\varphi_2, \varphi_n); \ldots & 2E_1(\varphi_n, \varphi_n) - \varkappa H_1(\varphi_n, \varphi_n) \end{vmatrix} = 0.$$

If we perform all the necessary calculations, then this equation assumes the following form:

$$\begin{vmatrix} \frac{(\lambda+2\mu)R^2h^3}{2} - \frac{R^2h^3}{24}\varkappa, & \frac{2\lambda R^4 h}{5}, & 0, & 0, & \frac{(\lambda+2\mu)R^2h^5}{2} - \frac{R^2h^2}{24}\varkappa \\ \frac{2\lambda R^4 h}{5}, & \frac{(74\lambda+124\mu)R^6h}{75} - \frac{3R^8h}{200}\varkappa, & 0, & 0, & \frac{2\lambda R^4h^3}{5} \\ 0, & 0, & \frac{6(\lambda+\mu)R^6}{25h} + \frac{\mu R^4 h}{5}\left(\frac{6R^2}{5h^2}+2\right)^2 - \left(\frac{3}{100}\frac{R^8}{h} + \frac{R^8h}{24}\right)\varkappa, & 0, & \frac{\lambda R^4 h}{5} \\ 0, & 0, & 0, & \frac{(\lambda+\mu)R^2h^5}{90} + \frac{\mu R^4h^3}{12} - \frac{R^4h^5}{720}\varkappa, & 0, \\ \frac{(\lambda+2\mu)R^2h^2}{2} - \frac{R^2h^4}{24}\varkappa, & \frac{2\lambda R^4h^2}{5}, & \frac{\lambda R^4 h}{5}, & 0, & \frac{2(\lambda+2\mu)R^2h^3}{3} - \frac{2R^2h^5}{45}\varkappa \end{vmatrix} = 0. \quad (15)$$

Equation (15) is degenerate, breaking down into two simpler equations since the fourth column contains only one non-zero term. One of these equations is

$$\frac{(\lambda+\mu)R^2h^5}{90}+\frac{\mu R^4h^3}{12}-\frac{R^4h^5}{720}\varkappa=0. \tag{16}$$

The second has the following form

$$\begin{vmatrix} \frac{(\lambda+2\mu)R^2h}{2}-\frac{R^2h^3}{24}\varkappa, & \frac{2\lambda R^4h}{5}, & 0, & \frac{(\lambda+2\mu)R^2h^2}{2}-\frac{R^2h^4}{24}\varkappa \\ \frac{2\lambda R^4h}{5}, & \frac{(74\lambda+124\mu)R^6h}{75}-\frac{3R^8h}{200}\varkappa, & 0, & \frac{2\lambda R^4h^2}{5} \\ 0, & 0, & \frac{6(\lambda+\mu)R^6}{25h}+\frac{\mu R^4h}{4}\left(\frac{6R^2}{5h^2}+2\right)^2-\left(\frac{3}{100}\frac{R^8}{h}+\frac{R^6h}{24}\right)\varkappa, & \frac{\lambda R^4h}{5} \\ \frac{(\lambda+2\mu)R^2h^2}{2}-\frac{R^2h^4}{24}\varkappa, & \frac{2\lambda R^4h^2}{5}, & \frac{\lambda R^4h^2}{5}, & \frac{2(\lambda+2\mu)R^2h^3}{3}-\frac{2R^2h^5}{45}\varkappa \end{vmatrix}=0. \tag{17}$$

In determinant (17) we multiply the first column by h and subtract from the fourth. Then the first two elements of the fourth column vanish. By the well known Laplace theorem the determinant in (17) breaks down into the product of two determinants and, correspondingly, equation (17) reduces to two quadratic equations

$$\begin{vmatrix} \frac{(\lambda+2\mu)R^2h}{2}-\frac{R^2h^3}{24}\varkappa, & \frac{2\lambda R^4h}{5} \\ \frac{2\lambda R^4h}{5}, & \frac{(74\lambda+124\mu)R^6h}{75}-\frac{3R^8h}{200}\varkappa \end{vmatrix}=0. \tag{18}$$

and

$$\begin{vmatrix} \frac{6(\lambda+\mu)R^6}{25h}+\frac{\mu R^4h}{4}\left(\frac{6R^2}{5h^2}+2\right)^2-\left(\frac{3R^8}{100h}+\frac{R^6h}{24}\right)\varkappa, & \frac{\lambda R^4h}{5} \\ \frac{\lambda R^4h}{5}, & \frac{(\lambda+2\mu)R^2h^3}{6}-\frac{R^2h^5}{360}\varkappa \end{vmatrix}=0. \tag{19}$$

The lowest of the roots of the equations (16), (18) and (19) gives (approximately) the value $\varkappa_1=\gamma\omega_1^2$, where, as we mentioned above, ω_1 denotes the lowest frequency of the eigenvibrations of the cylinder.

Suppose, for example that $\sigma=\frac{1}{3}$ or, what is the same thing, $\lambda=2\mu$. We put further $h=4$, $R=2$ and introduce the notation $\frac{\varkappa}{\mu}=\nu$. Solving equations (16), (18) and (19) for the values mentioned, we find that the lowest root $\nu_1=2\cdot722808$ and, consequently,

$$\omega_1\approx 0\cdot1650051\sqrt{\left(\frac{\mu}{\gamma}\right)}$$

If we equate the diagonal minors of determinant (15) to zero we obtain the following approximations (the superscript denotes the number of the approximation):

$$\nu_1^{(1)}=3,\quad \nu_1^{(2)}=\nu_1^{(3)}=\nu_1^{(4)}=\nu_1^{(5)}=2\cdot722808.$$

§ 71. Plane vibrations of an elastic rectangular plate

As a specific problem, we shall discuss a plate whose boundary is free from the influence of external forces. The equations for the free vibrations of an elastic plate have the form

$$\begin{aligned}-\left(\frac{\partial\sigma_x}{\partial x}+\frac{\partial\tau_{xy}}{\partial y}\right)-\omega^2\varrho u_x=0;\\ -\left(\frac{\partial\tau_{xy}}{\partial x}+\frac{\partial\sigma_y}{\partial y}\right)-\omega^2\varrho u_y=0,\end{aligned}\qquad(1)$$

where ϱ is the density of the elastic medium, λ and μ are its Lamé coefficients, and ω is the frequency of the free vibrations. According to § 39, the lowest eigenfrequency is determined from the relation

$$\varrho\omega_1^2=\min\frac{2E(\mathbf{u})}{H(\mathbf{u})}.\qquad(2)$$

We denote by $E(\mathbf{u})$ the integral of the elastic deformation energy, equal to

$$2E(\mathbf{u})=\int_{\Omega}\left\{(\lambda+2\mu)\theta^2-4\mu\frac{\partial u_x}{\partial x}\frac{\partial u_y}{\partial y}+\mu\left(\frac{\partial u_x}{\partial y}+\frac{\partial u_y}{\partial x}\right)^2\right\}\mathrm{d}\Omega, \quad (3)$$

$$\theta=\frac{\partial u_x}{\partial x}+\frac{\partial u_y}{\partial y}$$

and by $H(u)$, the integral

$$H(\mathbf{u})=\int_{\Omega}(u_x^2+u_y^2)\mathrm{d}\Omega. \quad (4)$$

The boundary condition of our problem is natural, therefore, in seeking the minimum of (2) we need not subject $\mathbf{u}$ to any boundary conditions; $\mathbf{u}$ however must be subject to the conditions which exclude rigid displacements of the plate viz.

$$\int_{\Omega}u_x\mathrm{d}\Omega=0, \quad \int_{\Omega}u_y\mathrm{d}\Omega=0, \quad \int_{\Omega}(xu_y-yu_x)\mathrm{d}\Omega=0. \quad (5)$$

For the sake of argument we shall consider the case of a rectangular plate.

We adopt the notation for the lengths of the sides and the disposition of the coordinate axes which was used in § 66. For the coordinate functions we take

$$\cos\frac{k\pi x}{a}\cos\frac{m\pi y}{b}, \quad k,\ m=0,\ 1,\ 2,\ \ldots\ .$$

Restricting ourselves to the values $k+m\leq 2$, we put

$$u_x\approx a_1\cos\frac{\pi x}{a}+a_2\cos\frac{2\pi x}{a}+a_3\cos\frac{\pi y}{b}+a_4\cos\frac{2\pi y}{b}$$

$$+a_5\cos\frac{\pi x}{a}\cos\frac{\pi y}{b},$$

$$u_y\approx b_1\cos\frac{\pi x}{a}+b_2\cos\frac{2\pi x}{a}+b_3\cos\frac{\pi y}{b}+b_4\cos\frac{2\pi y}{b}$$

$$+b_5\cos\frac{\pi x}{a}\cos\frac{\pi y}{b}. \quad (6)$$

By virtue of the third of equations (5), the coefficients a_k and b_k are connected by the relation

$$ab_1 = a_3 b. \tag{7}$$

The first two of equations (5) are satisfied automatically since the expressions (6) do not contain independent terms.

Formulae (6) are conveniently treated in the following fashion.

We introduce nine vectors φ_k, $k=1, 2, \ldots, 9$ by putting

$$\varphi_1 = \left(\cos\frac{\pi x}{a}, 0\right), \qquad \varphi_2 = \left(\cos\frac{\pi y}{b}, \frac{b}{a}\cos\frac{\pi x}{a}\right),$$

$$\varphi_3 = \left(\cos\frac{2\pi x}{a}, 0\right), \qquad \varphi_4 = \left(0, \cos\frac{2\pi x}{a}\right),$$

$$\varphi_5 = \left(0, \cos\frac{\pi y}{b}\right), \qquad \varphi_6 = \left(\cos\frac{2\pi y}{b}, 0\right),$$

$$\varphi_7 = \left(0, \cos\frac{2\pi y}{b}\right), \qquad \varphi_8 = \left(\cos\frac{\pi x}{a}\cos\frac{\pi y}{b}, 0\right),$$

$$\varphi_9 = \left(0, \cos\frac{\pi x}{a}\cos\frac{\pi y}{b}\right).$$

Formulae (6) and (7) show that the required vector $\mathbf{u}$ can be expressed approximately in terms of linear combinations of the vectors $\varphi_1, \varphi_2, \ldots, \varphi_9$.

We use the notation $\dfrac{ab\varrho\omega^2}{\mu} = \varkappa$. In order to determine $\varkappa$, we use equation (4) § 33. In our case when each row of the determinant has been divided by $\dfrac{\mu}{2}$, it assumes the form (8), page 443, where σ is Poisson's ratio.

We note that equation (8) can be simplified since the second column of determinant (8) contains only one non-zero element, and that it breaks down into two simpler equations of which the first is obtained by equating the element men-

$$
\begin{vmatrix}
\frac{2\pi^2(1-\sigma)}{1-2\sigma}\frac{b}{a}-\varkappa, & 0 & 0 & 0 & \frac{16\sigma}{1-2\sigma} & 0 & 0 & 0 & 0 \\
0, & \pi^2\left(\frac{a}{b}+\frac{b^3}{a^3}+\frac{16}{\pi^2}\frac{b}{a}\right)-\varkappa\frac{a+b}{a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0, & \frac{8\pi^2(1-\sigma)}{1-2\sigma}\frac{b}{a}-\varkappa & 0 & 0 & 0 & 0 & 0 & \frac{32\sigma}{3(1-2\sigma)} \\
0 & 0 & 0 & 4\pi^2\frac{b}{a}-\varkappa & 0 & 0 & 0 & \frac{32}{3} & 0 \\
\frac{16\varrho}{1-2\sigma} & 0 & 0 & 0, & \frac{2\pi^2(1-\sigma)}{1-2\sigma}\frac{a}{b}-\varkappa, & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 4\pi^2\frac{a}{b}-\varkappa & 0 & 0 & \frac{32}{3} \\
0 & 0 & 0 & 0 & 0 & 0, & \frac{8\pi^2(1-\sigma)}{1-2\sigma}\frac{a}{b}-\varkappa, & \frac{64(1-\sigma)}{3(1-2\sigma)} & 0 \\
0 & 0 & 0 & \frac{32}{3} & 0 & 0 & \frac{64(1-\sigma)}{3(1-2\sigma)}, & \frac{\pi^2}{2}\left[\frac{2(1-\sigma)}{1-2\sigma}\frac{b}{a}+\frac{a}{b}\right]-\frac{\varkappa}{2} & 0 \\
0 & 0 & \frac{32\sigma}{1-2\sigma} & 0 & 0 & \frac{32}{3} & 0 & 0, & \frac{\pi^2}{2}\left[\frac{2(1-\sigma)}{1-2\sigma}\frac{a}{b}+\frac{b}{a}\right]-\frac{\varkappa}{2}
\end{vmatrix}
=0. \quad (8)
$$

tioned to zero, and the second, by equating the minor of this element to zero. The first equation has the root

$$\varkappa=\frac{a\pi^2}{a+b}\left(\frac{a}{b}+\frac{b^3}{a^3}+\frac{16}{\pi^2}\frac{b}{a}\right). \tag{9}$$

In order to solve the second equation we give numerical values for σ and $\frac{b}{a}$. Suppose, for example, that $\sigma=\frac{1}{3}$, $\frac{b}{a}=\frac{1}{2}$. Then this equation assumes the form

$$\begin{vmatrix}
2\pi^2-\varkappa & 0 & 0 & 16 & 0 & 0 & 0 & 0 \\
0 & 8\pi^2-\varkappa & 0 & 0 & 0 & 0 & 0 & \frac{64}{3} \\
0 & 0 & 2\pi^2-\varkappa & 0 & 0 & 0 & \frac{32}{3} & 0 \\
16 & 0 & 0 & 8\pi^2-\varkappa & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 8\pi^2-\varkappa & 0 & 0 & \frac{32}{3} \\
0 & 0 & 0 & 0 & 0 & 32\pi^2-\varkappa & \frac{64}{3} & 0 \\
0 & 0 & \frac{64}{3} & 0 & 0 & \frac{128}{3} & 4\pi^2-\varkappa & 0 \\
0 & \frac{128}{3} & 0 & 0 & \frac{64}{3} & 0 & 0 & \frac{17\pi^2}{2}-\varkappa
\end{vmatrix}=0. \tag{10}$$

We note that, for $\sigma=\frac{1}{3}$ and $\frac{b}{a}=\frac{1}{2}$, the quantity (9) equals $\varkappa_0=19{\cdot}315273$.

In order to follow the process, we shall find the approximate values of the lowest root $\varkappa_1$ by equating the diagonal minors of determinant (10) to zero. Equating the minor of the first three orders to zero we obtain values of $\varkappa_1$ equal to 19·739209 and 78·956835. Comparing this with $\varkappa_0$ we obtain as an approximate value of the lowest root of equation (8) the value 19·315273.

We now equate the fourth-order minor to zero.

$$\begin{vmatrix}
2\pi^2-\varkappa & 0 & 0 & 16 \\
0 & 8\pi^2-\varkappa & 0 & 0 \\
0 & 0 & 2\pi^2-\varkappa & 0 \\
16 & 0 & 0 & 8\pi^2-\varkappa
\end{vmatrix}=0.$$

The lowest root of this equation equals 15·692682. Since this number is less than $\varkappa_0$, it must be considered a more precise approximate value for the quantity $\frac{ab\varrho\omega_1^2}{\mu}$ where ω_1 denotes the lowest eigenfrequency of the vibrations of the plate.

By equating higher order minors to zero we obtain the lowest root of 15·692682 each time. Thus, to within the accuracy limits that we have adopted,

$$\frac{ab\varrho\omega_1^2}{\mu}=15\cdot692682.$$

72. Stability of a compressed elliptic plate

We consider an elliptic plate with semi-axes a and b $(a>b)$ the rim of which is fixed. We suppose that the plate is subjected to a uniform pressure $\sigma_x=\sigma_y=-\frac{\lambda D}{h}$, $\tau_{xy}=0$. We shall find the lowest value of the load at which the plate loses its stability.

Equation (4) § 38 in our case assumes the form

$$\nabla^4 w+\lambda\nabla^2 w=0. \tag{1}$$

The problem consists of finding the least eigenvalue of equation (1) under the boundary conditions

$$w=0, \quad \frac{\partial w}{\partial n}=0. \tag{2}$$

In applying the Ritz method to our problem we take as the coordinate functions

$$x^m y^n\left(1-\frac{x^2}{a^2}-\frac{y^2}{b^2}\right)^2, \quad m,\ n=0,\ 1,\ 2,\ \ldots \tag{3}$$

(as usual, we take axes of coordinates along the axes of symmetry of the plate).

Restricting ourselves to the values $m+n\leq 2$ we obtain

$$w\approx\left(1-\frac{x^2}{a^2}-\frac{y^2}{b^2}\right)^2(a_1+a_2x+a_3y+a_4x^2+a_5xy+a_6y^2). \qquad (4)$$

We put $\frac{a}{b}=\gamma$, $\lambda a^2=\varkappa$.

As usual, the Ritz method leads us to the equation (we omit the details of the calculations)

$$\begin{vmatrix}
\frac{8}{\gamma}(3+2\gamma^2+3\gamma^4)-\frac{2}{\gamma}(1+\gamma^2)\varkappa & 0 & 0, \frac{1}{3}\left(\frac{3}{\gamma}+2\gamma+3\gamma^3\right)-\frac{\gamma}{15}\varkappa & 0, \frac{1}{3}\left(\frac{3}{\gamma}+2\gamma+3\gamma^3\right)-\frac{1}{15\gamma}\varkappa \\
0 & 2\gamma+\frac{5}{\gamma}+\gamma^3-\frac{1}{5}\times\left(\frac{1}{\gamma}+\frac{\gamma}{3}\right)\varkappa & 0 & 0 & 0 \\
0 & 0 & \frac{2}{\gamma}+5\gamma+\frac{1}{\gamma^3}-\frac{1}{5}\left(\frac{1}{3\gamma^3}+\frac{1}{\gamma}\right)\varkappa & 0 & 0 \\
\gamma\left(\frac{3}{\gamma^2}+2+3\gamma^2\right)-\frac{\gamma}{5}\varkappa & 0 & 0 & \frac{27}{10\gamma}+\frac{3\gamma^3}{10}+\frac{11\gamma}{15}-\frac{1}{20\gamma}\left(1+\frac{\gamma^3}{3}\right)\varkappa & 0 \\
0 & 0 & 0 & 0 & \frac{5}{\gamma^3}+5\gamma^3+6-\frac{1}{6}\left(\frac{1}{\gamma^3}+1\right)\varkappa \\
\frac{1}{\gamma}\left(\frac{3}{\gamma^2}+2+3\gamma^2\right)-\frac{1}{5\gamma^3}\varkappa & 0 & 0 & \frac{1}{10\gamma^3}+\frac{1}{3\gamma}+\frac{\gamma}{10} & 0, \frac{27\gamma}{10}+\frac{3}{10\gamma^3}+\frac{11}{15\gamma}-\frac{1}{20\gamma}\left(+1\frac{1}{3\gamma^3}\right)\varkappa
\end{vmatrix}=0. \qquad (5)$$

The second, third and fifth columns of determinant (5) contain only one non-zero element each. Consequently equation (5) breaks down into the following four equations.

$$2\gamma+\frac{5}{\gamma}+\gamma^3-\frac{1}{5}\left(\frac{1}{\gamma}+\frac{\gamma}{3}\right)\varkappa=0; \qquad (6)$$

$$\frac{2}{\gamma}+5\gamma+\frac{1}{\gamma^3}-\frac{1}{5}\left(\frac{1}{3\gamma^3}+\frac{1}{\gamma}\right)\varkappa=0; \qquad (7)$$

$$5\left(\frac{1}{\gamma^2}+\gamma^2\right)+6-\frac{1}{6}\left(\frac{1}{\gamma^2}+1\right)\varkappa=0; \qquad (8)$$

$$\begin{vmatrix} \frac{8}{\gamma}(3+2\gamma^2+3\gamma^4) & \frac{1}{3}\left(\frac{3}{\gamma}+2\gamma+3\gamma^2\right) & \frac{1}{3}\left(\frac{3}{\gamma}+2\gamma+3\gamma^2\right) \\ -\frac{2}{\gamma}(1+\gamma^2)\varkappa, & -\frac{\gamma}{15}\varkappa, & -\frac{1}{15\gamma}\varkappa \\ \gamma\left(\frac{3}{\gamma^2}+2+3\gamma^2\right) & \frac{27}{10\gamma}+\frac{3\gamma^2}{10}+\frac{11\gamma}{15} & \frac{1}{10\gamma}+\frac{\gamma}{3}+\frac{\gamma^3}{10} \\ -\frac{\gamma}{5}\varkappa, & -\frac{1}{20\gamma}\left(1+\frac{\gamma^2}{3}\right)\varkappa, & \\ \frac{1}{\gamma}\left(\frac{3}{\gamma^2}+2+3\gamma^2\right) & \frac{1}{10\gamma^3}+\frac{1}{3\gamma}+\frac{\gamma}{10}, & \frac{27\gamma}{10}+\frac{3}{10\gamma^3}+\frac{11}{15\gamma} \\ -\frac{1}{5\gamma^3}\varkappa & & -\frac{1}{20\gamma}\left(1+\frac{1}{3\gamma^2}\right)\varkappa \end{vmatrix} = 0. \quad (9)$$

If $\varkappa_1$ is the least of the roots of equations (6)–(9), then the approximate value of the critical load equals $\frac{D\varkappa_1}{a^2h}$.

As an example we take $\gamma=2$. The roots of equations (6), (7) and (8) equal 78; 102·69; 121·8 respectively. Equation (9) for $\gamma=2$ assumes the form

$$\begin{vmatrix} 472-10\varkappa & \frac{59}{3}-\frac{4}{15}\varkappa & \frac{59}{3}-\frac{1}{15}\varkappa \\ 59-\frac{4}{5}\varkappa & \frac{313}{30}-\frac{7}{60}\varkappa & \frac{91}{30} \\ \frac{59}{4}-\frac{1}{20}\varkappa & \frac{91}{120} & \frac{1393}{120}-\frac{13}{240}\varkappa \end{vmatrix} = 0,$$

or, expanding the determinant and dividing by the coefficient of $\varkappa^3$,

$$\varkappa^3-343\cdot5707\varkappa^2+32028\cdot293\varkappa-806705\cdot6=0.$$

We solve this equation by Newton's method taking the root of the first-order diagonal minor, $\varkappa'=47\cdot2$, as the first approximation. The second approximation given by Newton's method will be $\varkappa''=42\cdot06696$, the third approximation is $\varkappa'''=42\cdot06681$. Thus to within a fairly high degree of precision, we can consider that $\varkappa_1=42\cdot067$.

CHAPTER IX

THE BUBNOV–GALERKIN METHOD

§ 73. Foundations of the method

The Bubnov–Galerkin method can be considered as a generalization of the Ritz method for an equation of the form $Au=f$, where the operator A is assumed to be positive-definite. Before formulating the method in its general form, we shall demonstrate it for a simple special case which, however, embraces an overwhelming majority of its applications.

Let the unknown function $u(P)$ satisfy the non-homogeneous equation

$$Lu-f(P)=0 \tag{1}$$

within some domain Ω and also, perhaps, certain *homogeneous* boundary conditions. We choose an infinite sequence of coordinate functions $\varphi_1(P)$, $\varphi_2(P)$, ..., $\varphi_n(P)$, ..., which are continuously differentiable in the closed domain $\bar{\Omega}=\Omega+S$ as many times as are requisite for the specified problem, and which satisfy *all* the boundary conditions of our problem. As usual we denote the boundary of the domain by S. We shall assume that both equation (1) and the boundary conditions corresponding to it are linear then. The function

$$u_n(P)=\sum_{k=1}^{n} a_k\varphi_k(P), \tag{2}$$

where a_k are arbitrarily chosen constants, then satisfies all the boundary conditions of the problem. *According to the*

Bubnov–Galerkin method the coefficients a_k *are determined from the requirement that the left-hand side of equation* (1) *after substituting* $u_n(P)$ *for* $u(P)$ *therein, is orthogonal to the functions* $\varphi_1(P), \varphi_2(P), \ldots, \varphi_n(P)$.

The Bubnov–Galerkin method thus leads to the following system of linear algebraic equations:

$$\sum_{k=1}^{n} a_k(L\varphi_k, \varphi_m) = (f, \varphi_m), \quad m = 1, 2, \ldots, n. \tag{3}$$

If the problem of the eigenvalues of the equation

$$Lu - \lambda u = 0,$$

is posed, then, in precisely the same way the Bubnov–Galerkin method leads to the system of equations

$$\sum_{k=1}^{n} a_k\{(L\varphi_k, \varphi_m) - \lambda(\varphi_k, \varphi_m)\} = 0, \qquad k = 1, 2, \ldots, n.$$

By equating the determinant of this system to zero, we obtain an equation which gives approximate values for the eigenvalues:

$$\begin{vmatrix} (L\varphi_1, \varphi_1) - \lambda(\varphi_1, \varphi_1); & (L\varphi_1, \varphi_2) - \lambda(\varphi_1, \varphi_2); \ldots; & (L\varphi_1, \varphi_n) - \lambda(\varphi_1, \varphi_n) \\ (L\varphi_2, \varphi_1) - \lambda(\varphi_2, \varphi_1); & (L\varphi_2, \varphi_2) - \lambda(\varphi_2, \varphi_2); \ldots; & (L\varphi_2, \varphi_n) - \lambda(\varphi_2, \varphi_n) \\ \ldots\ldots\ldots\

We construct an approximate solution of equation (4) in the form of a linear combination of the coordinate elements with constant coefficients (2). We determine the coefficients a_k from the condition that, after replacing u by u_n, the left-hand side of equation (4) is orthogonal to the elements $\varphi_1, \varphi_2, \ldots, \varphi_n$. This leads to the following system of linear equations in the unknowns a_k:

$$\sum_{k=1}^{n}(A\varphi_k, \varphi_j)a_k=(f, \varphi_j), \qquad j=1, 2, \ldots, n. \tag{5}$$

System (5) is identical in form to system (8) § 14 to which the Ritz method leads. Hence it is readily concluded that the Bubnov–Galerkin and Ritz methods are identical if operator A is positive-bounded-below. In the more general case where A is only positive definite the Ritz method is inapplicable whereas the Bubnov–Galerkin method can be used.

Of great importance is the question of the convergence of the Bubnov–Galerkin method. The history of this question has been given in detail in the introduction. In the next two sections we shall give a comparatively elementary proof (based on certain theorems of the theory of integral equations) of the convergence for those cases where equation (4) is either a Fredholm-type integral equation (the first proof for which was given by Yu. V. Repman [1]) or an ordinary second-order differential equation. In the following sections we shall give a fairly general sufficiency criterion for the convergence of the Bubnov–Galerkin method which we shall apply to partial differential equations. We shall consider in particular the question of natural boundary conditions.

§ 74. Proof of convergence for Fredholm integral equations

Suppose we are given the integral equation

$$u(P)-\int_{\Omega}K(P, Q)u(Q)\mathrm{d}\Omega_Q-f(P)=0, \tag{1}$$

concerning which we assume the following:
1)

$$\int_\Omega \int_\Omega K^2(P, Q)\mathrm{d}\Omega_P \mathrm{d}\Omega_Q < \infty; \quad \int_\Omega f^2(P)\mathrm{d}\Omega < \infty;$$

2) equation (1) has a solution which is unique and satisfies the inequality

$$\|u\|^2 = \int_\Omega u^2(P)\mathrm{d}\Omega < \infty. \tag{2}$$

We now apply the Bubnov–Galerkin method to our equation. Firstly, we subject the coordinate functions $\varphi_k(P)$, $k=1, 2, \ldots$ to the process of orthogonalization (if necessary); it is readily seen that the approximate solutions $u_n(P)$ are unchanged by this. Now the φ_k satisfy the equations

$$(\varphi_j, \varphi_k) = \int_\Omega \varphi_j(P)\varphi_k(P)\mathrm{d}\Omega = \begin{cases} 1, & j=k, \\ 0, & j \neq k. \end{cases} \tag{3}$$

We require the coordinate functions to be complete in the sense of convergence in the mean. This means that for any function $u(P)$ with finite norm and for any positive number ε, it is possible to find a positive integer N and constants $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that

$$\left\| u - \sum_{k=1}^{n} \alpha_k \varphi_k \right\|^2 = \int_\Omega \left[u(P) - \sum_{k=1}^{n} \alpha_k \varphi_k(P) \right]^2 \mathrm{d}\Omega < \varepsilon^2.$$

We now put

$$u_n(P) = \sum_{k=1}^{n} a_k \varphi_k(P),$$

we substitute this for $u(P)$ in the left-hand side of equation (1) and we require that the result of the substitution be orthogonal to the functions $\varphi_1(P), \varphi_2(P), \ldots, \varphi_n(P)$. Using equations (3) we obtain a system of linear algebraic equations in the unknowns $a_1, a_2, \ldots, a_n$:

$$a_m - \sum_{k=1}^{n} \gamma_{mk} a_k = (f, \varphi_m), \qquad m = 1, 2, \ldots, n, \tag{4}$$

where

$$\gamma_{mk}=\int_{\Omega}\int_{\Omega}K(P, Q)\varphi_m(P)\varphi_k(Q)\mathrm{d}\Omega_P\mathrm{d}\Omega_Q.$$

The numbers γ_{mk} are the Fourier coefficients of the function $K(P, Q)$ when it is expanded into a double Fourier series in the orthogonal functions φ_k; in exactly the same way the (f, φ_m) are the Fourier coefficients of the function $f(P)$.

We now introduce the notation:

$$K_n(P, Q)=\sum_{k, m=1}^{n}\gamma_{mk}\varphi_m(P)\varphi_n(Q),$$

$$f_n(P)=\sum_{m=1}^{n}(f, \varphi_m)\varphi_m(P).$$

The system of orthogonal and normalized functions $\varphi_m(P)$ is complete, therefore

$$\left.\begin{aligned}&\lim_{n\to\infty}\int_{\Omega}\int_{\Omega}[K_n(P, Q)-K(P, Q)]^2\mathrm{d}\Omega_P\mathrm{d}\Omega_Q=0,\\&\lim_{n\to\infty}\int_{\Omega}[f_n(P)-f(P)]^2\mathrm{d}\Omega=\lim_{n\to\infty}\|f_n-f\|^2=0.\end{aligned}\right\}\tag{5}$$

We consider the auxiliary integral equation

$$v_n(P)-\int_{\Omega}K_n(P, Q)v_n(Q)\mathrm{d}\Omega_Q=f_n(P).\tag{6}$$

The remainder of the proof is based upon a theorem from the theory of integral equations. By virtue of this it follows from relations (5) that equation (6) is soluble for sufficiently large n and has a unique solution provided that equation (1) possesses the given properties and that, in the mean

$$v_n(P)\xrightarrow[n\to\infty]{}v(P).\tag{7}$$

To avoid interrupting the argument we shall give the proof of this theorem at the end of the section.

Integral equation (6) is readily solved. Substituting for $K_n(P, Q)$ and $f_n(P)$ in (6), we find that

$$v_n(P)=\sum_{k=1}^{m}A_k\varphi_k(P),\tag{8}$$

where

$$A_m = \sum_{k=1}^{m} \gamma_{mk} \int_{\Omega} v_n(Q)\varphi_k(Q)\mathrm{d}\Omega + (f, \varphi_m).$$

Substituting the value of $v_n(Q)$ from (8) and using inequalities (3), we find that the A_k satisfy the system

$$A_m - \sum_{k=1}^{n} \gamma_{mk} A_k = (f, \varphi_m), \quad m = 1, 2, \ldots, n, \tag{9}$$

which is identical to system (4). For n sufficiently large, system (9) and also equation (6) have a solution which is unique, therefore $A_k = a_k$, $k = 1, 2, \ldots, n$ and

$$v_n(P) \equiv u_n(P).$$

Now from (7) it follows that, in the mean

$$u_n(P) \xrightarrow[n\to\infty]{} u(P).$$

Thus the approximate solution of a Fredholm integral equation constructed by the Bubnov-Galerkin method converges in the mean to the exact solution of this equation, if the system of coordinate functions is complete in the sense of convergence in the mean.

We now give a proof of the theorem mentioned previously in this section.

It has been assumed that equation (1) is soluble and has a unique solution. This means that unity is not a characteristic value for the kernel and that there exists a resolvent $\Gamma(P, Q)$ with the aid of which the solution of equation (1) may be represented in the form

$$u(P) = f(P) + \int_{\Omega} \Gamma(P, Q) f(Q)\mathrm{d}\Omega_Q. \tag{10}$$

The resolvent itself satisfies an inequality of the form:

$$\int_{\Omega}\int_{\Omega} \Gamma^2(P, Q)\,\mathrm{d}\Omega_P \mathrm{d}\Omega_Q < \infty. \tag{11}$$

We write down equation (6) in the form

$$v_n(P) - \int_{\Omega} K(P, Q)\, v_n(Q)\,\mathrm{d}\Omega_Q$$
$$= f_n(P) - \int_{\Omega} [K(P, Q) - K_n(P, Q)]\, v_n(Q)\,\mathrm{d}\Omega_Q. \tag{12}$$

Considering for the moment that the right-hand side of the last equation is the unknown, we can use formula (10). This leads to an integral equation equivalent to equation (6):

$$v_n(P) = \int_{\Omega} \widetilde{K}_n(P, Q)\, v_n(Q)\, d\Omega_Q + f_n(P) + \int_{\Omega} \Gamma(P, Q)\, f_n(Q)\, d\Omega_Q. \tag{13}$$

Here for brevity we have put

$$\widetilde{K}_n(P, Q) = K_n(P, Q) - K(P, Q) + \int_{\Omega} \Gamma(P, R)\,[K_n(R, Q) - K(R, Q)]\, d\Omega_R. \tag{14}$$

We now consider the integral of the square of the kernel $\widetilde{K}_n$. Firstly, by Cauchy's inequality

$$\widetilde{K}_n^2(P, Q) \leq [K_n(P, Q) - K(P, Q)]^2 + 2\int_{\Omega} \Gamma^2(P, R)\, d\Omega_R \int_{\Omega} [K_n(R, Q) - K(R, Q)]^2 d\Omega_R.$$

Integrating, we have

$$\int_{\Omega}\int_{\Omega} \widetilde{K}_n^2\, d\Omega_P d\Omega_Q \leq 2\int_{\Omega}\int_{\Omega} [K_n(P, Q) - K(P, Q)]^2 d\Omega_P d\Omega_Q + 2\int_{\Omega}\int_{\Omega} \Gamma^2(P, R)\, d\Omega_P d\Omega_R \times \int_{\Omega}\int_{\Omega} [K_n(P, R) - K(P, R)]^2 d\Omega_P d\Omega_R. \tag{15}$$

In inequality (15) the right-hand side tends to zero as $n \to \infty$. Consequently there exists an N such that, for $n \geq N$

$$\int_{\Omega}\int_{\Omega} \widetilde{K}_n^2(R, Q)\, d\Omega_P d\Omega_Q < 1.$$

Thus for $n \geq N$ equation (13) and equation (6) have the same solution. This solution can be constructed by the method of successive approximations (cf. the author's book [7], § 2).

We now estimate difference $u(P) - v_n(P)$. Subtracting equations (1) and (12) term by term, we obtain

$$u(P) - v_n(P) - \int_{\Omega} K(P, Q)\,[u(Q) - v_n(Q)]\, d\Omega_Q = f(P) - f_n(P) + \int_{\Omega} [K(P, Q) - K_n(P, Q)]\, v_n(Q)\, d\Omega_Q. \tag{16}$$

Denoting the right-hand side of the last equation by $F_n(P)$ for brevity, we have by formula (10)

$$u(P)-v_n(P)=F_n(P)+\int_\Omega \Gamma(P, Q)\,F_n(Q)\,d\Omega_Q.$$

Hence

$$\|u-v_n\|\le\|F_n\|+\left\|\int_\Omega \Gamma(P, Q)\,F_n(Q)\,d\Omega_Q\right\|.$$

By Cauchy's inequality

$$\left(\int_\Omega \Gamma(P, Q)\,F_n(Q)\,d\Omega_Q\right)^2\le\int_\Omega \Gamma^2(P, Q)d\Omega_Q\cdot\|F_n\|^2.$$

Integrating this over Ω, we find

$$\left\|\int_\Omega \Gamma(P, Q)\,F_n(Q)\,d\Omega_Q\right\|^2\le\|F_n\|^2\int_\Omega\int_\Omega \Gamma^2(P, Q)d\Omega_P d\Omega_Q$$

and, consequently,

$$\|u-v_n\|\le C\|F_n\|,\quad C=1+\left\{\int_\Omega\int_\Omega \Gamma^2(P, Q)d\Omega_P d\Omega_Q\right\}^{\frac{1}{2}}. \tag{17}$$

For brevity we use the notation

$$\varepsilon_n^2=\int_\Omega\int_\Omega [K(P, Q)-K_n(P, Q)]^2 d\Omega_P d\Omega_Q.$$

Obviously $\varepsilon_n\to 0$ as $n\to\infty$. We shall consider n to be so great that, in every case, $\varepsilon_n<\frac{1}{2C}$. Applying Cauchy's inequality and the triangle inequality for the norm, we readily find

$$\|F_n\|\le\|f-f_n\|+\varepsilon_n\|v_n\|,$$

or

$$\|F_n\|\le\|f-f_n\|+\varepsilon_n\|u-(u-v_n)\|\le\|f-f_n\|+\varepsilon_n\|u\|+\varepsilon_n\|u-v_n\|.$$

Substituting this in (17) we obtain

$$(1-C\varepsilon_n)\|u-v_n\|\le C\|f-f_n\|+C\varepsilon_n\|u\|.$$

We may strengthen the inequality by replacing $1-C\varepsilon_n$ on the left by $\frac{1}{2}$. This gives

$$\|u-v_n\|<2C\|f-f_n\|+C\varepsilon_n\|u_n\|\underset{n\to\infty}{\to}0.$$

§ 75. Proof of convergence for ordinary second-order differential equations

We consider the problem of integrating the equation

$$-y''+p(x)y'+q(x)y=f(x) \tag{1}$$

under the boundary conditions

$$y(0)=y(1)=0. \tag{2}$$

For simplicity we shall consider that the coefficients $p(x)$ and $q(x)$ are continuous in the range $0 \leq x \leq 1$. We further suppose that the function $f(x)$ has a finite norm. We shall assume that the boundary value problem that we have posed has a unique solution. We select a system of coordinate functions $\varphi_n(x)$, $n = 1, 2, \ldots$ which are continuously differentiable and satisfy boundary conditions (2), and we require that this system be complete in the following sense: *the derivative* of any function which is continuous and continuously differentiable for $0 \leq x \leq 1$ can be approximated in the mean as closely as required by linear combinations *of the derivatives* $\varphi_n'(x)$. We put

$$y_n = \sum_{k=1}^{n} a_k \varphi_k(x).$$

The Bubnov–Galerkin system of equations has the form

$$\sum_{k=1}^{n} a_k \left\{ -\int_0^1 \varphi_k'' \varphi_m \mathrm{d}x + \int_0^1 (p\varphi_k' + q\varphi_k)\varphi_m \mathrm{d}x \right\} = \int_0^1 f\varphi_m \mathrm{d}x,$$

$$m = 1, 2, \ldots, n.$$

We integrate the first integral on the left by parts; and using condition (2) we obtain

$$\sum_{k=1}^{n} a_k \int_0^1 (\varphi_k'\varphi_m' + p\varphi_k'\varphi_m + q\varphi_k\varphi_m)\mathrm{d}x = \int_0^1 f\varphi_m \mathrm{d}x, \tag{3}$$

$$m = 1, 2, \ldots, n.$$

System (3) can be simplified somewhat if the derivatives $\varphi_k'(x)$ are subjected to the process of orthogonalization. In this case

$$\int_0^1 \varphi_k'\varphi_m' \mathrm{d}x = \begin{cases} 1, & k = m, \\ 0, & k \neq m, \end{cases}$$

and system (3) assumes the form

$$a_m + \sum_{k=1}^{n} a_k \int_0^1 (p\varphi_k'\varphi_m + q\varphi_k\varphi_m)\,\mathrm{d}x = \int_0^1 f\varphi_m \mathrm{d}x, \tag{4}$$

$$m = 1, 2, \ldots n.$$

We now replace $\varphi_k(x)$ and $\varphi_m(x)$ by their expressions in terms of the derivatives $\varphi_k'(x)$ and $\varphi_m'(x)$. Since $\varphi_k(0)=0$

$$\varphi_k(x)=\int_0^x \varphi_k'(s)\mathrm{d}s.$$

Substituting this in (4), we obtain after simple manipulation

$$a_m+\sum_{k=1}^{n} a_k \int_0^1 \varphi_m'(x)\mathrm{d}x \int_0^1 \left[p(s)\varphi_m'(s)+q(s)\int_0^s \varphi_k'(t)\,\mathrm{d}t\right]\mathrm{d}s$$

$$=\int_0^1 f_1(x)\varphi_k'(x)\mathrm{d}x, \tag{5}$$

where

$$f_1(x)=\int_0^x f(t)\mathrm{d}t.$$

Equation (1) under boundary conditions (2) can be reduced to an equivalent integral equation. Let

$$G(x,\,s)=\begin{cases} x(1-s), & x\le s,\\ s(1-x), & s\le x.\end{cases}$$

Then, if $y(x)$ satisfies conditions (2), it is comparatively easy to verify the equation

$$y(x)=-\int_0^1 G(x,\,s)y''(s)\mathrm{d}s.$$

Substituting the value of y'' from equation (1), we obtain

$$y(x)=-\int_0^1 G(x,\,s)\{p(s)y'(s)+q(s)y(s)-f(s)\}\mathrm{d}s.$$

We differentiate this equation with respect to x and replace $y(s)$ in the integral by

$$\int_0^s y'(t)\mathrm{d}t.$$

Performing some simple manipulation, we obtain an integral equation in $y'(x)$

$$y'(x)=\int_0^1 s\left[p(s)y'(s)+q(s)\int_0^s y'(t)\mathrm{d}t-f(s)\right]\mathrm{d}s$$
$$-\int_x^1\left[p(s)y'(s)+q(s)\int_0^s y'(t)\mathrm{d}t\right]\mathrm{d}s+f_1(x). \qquad (6)$$

We now apply the Bubnov–Galerkin method to this equation. We take the functions $\varphi_k'(x)$ as the coordinate functions. We put

$$y_n'(x)=\sum_{k=1}^n a_k\varphi_k(x).$$

We substitute this in (6) and carry all terms over to the left. We also require that the expression obtained is orthogonal to $\varphi_1'(x)$, $\varphi_2'(x)$, ..., $\varphi_n'(x)$. We then obtain n equations in the a_k. These equations as is easily seen, are the same as equations (5). Thus the application of the Bubnov–Galerkin method to the differential equation (1) is equivalent to the application of the same method to the integral equation (6). As was established in the previous section, the approximate solutions $y_n'(x)$ for this equation converge in the mean to $y'(x)$, so that

$$\lim_{n\to\infty}\int_0^1[y_n'(x)-y'(x)]^2\mathrm{d}x=0.$$

Hence it readily follows that $y_n(x)\to y(x)$ uniformly. In fact since

$$y_n(x)-y(x)=\int_0^x[y_n'(s)-y'(s)]\mathrm{d}s.$$

Hence by Cauchy's inequality

$$|y_n(x)-y(x)|$$
$$\leq\sqrt{(x)}\sqrt{\left(\int_0^x[y_n'(s)-y'(s)]^2\,\mathrm{d}s\right)}\leq\sqrt{\left(\int_0^1[y_n'(s)-y'(s)]^2\mathrm{d}s.\right)}$$

The last integral does not depend upon x and tends to zero as $n \to \infty$; from Weierstrass's theorem $y_n(x) \to y(x)$ uniformly.

Thus we have established the convergence of the Bubnov–Galerkin method for the problem considered.

§ 76. Completely continuous operators

An operator A is called *degenerate* or *finite* if it can be represented in the form

$$Au = \sum_{k=1}^{n} (u, \psi_k)\varphi_k, \tag{1}$$

where the number n is finite and φ_k and ψ_k are given elements of the Hilbert space considered. An operator A is called *completely continuous* if, for any $\varepsilon > 0$, it can be represented in the form

$$Au = A'_\varepsilon u + A''_\varepsilon u, \tag{2}$$

where A'_ε is a degenerate operator and the norm of the operator A''_ε can be made smaller than any number $\varepsilon > 0$ specified beforehand. (Usually another definition of a completely continuous operator, which is equivalent to that given (cf. theorem 1 below) is adopted.)

We note some properties of completely continuous operators.

a) *Any degenerate operator is completely continuous.* In fact the degenerate operator can be represented in form (2)—it suffices to put $A''_\varepsilon = 0$.

b) *A completely continuous operator is bounded.*

Let $A'_\varepsilon u = \sum_{k=1}^{n} (u, \psi_k)\varphi_k$ and let $\|A''_\varepsilon\| < \varepsilon$. Then

$$\|Au\| \leq \sum_{k=1}^{n} |(u, \psi_k)| \, \|\varphi_k\| + \varepsilon\|u\| \leq \left\{ \sum_{k=1}^{n} \|\psi_k\| \, \|\varphi_k\| + \varepsilon \right\} \|u\|.$$

Hence it is clear that operator A is bounded and

$$\|A\| \leq \sum_{k=1}^{n} \|\psi_k\| \, \|\varphi_k\| + \varepsilon.$$

c) *The sum of a finite number of completely continuous operators is completely continuous.* The proof is obvious.

d) *If operator A is bounded and operator B is completely continuous, then the operators AB and BA are completely continuous.* We specify an arbitrary number $\varepsilon>0$ and suppose that

$$Bu=\sum_{k=1}^{n}(u,\ \psi_k)\varphi_k+B''u,\ \ \|B''\|<\frac{\varepsilon}{\|A\|}.$$

Then $BAu=\sum_{k=1}^{n}(Au,\ \psi_k)\,\varphi_k+B''Au$. It is clear from the inequality $|(Au,\ \psi_k)|\leq\|\varphi_k\|\,\|Au\|\leq\|\psi_k\|\,\|A\|\,\|u\|$ that $(Au,\ \varphi_k)$ is a bounded functional of u. By theorem 1 § 45 there exists an element $\tilde{\psi}_k$ such that $(Au,\ \psi_k)=(u,\ \tilde{\psi}_k)$.

Now

$$BAu=\sum_{k=1}^{n}(u,\ \tilde{\psi}_k)\varphi_k+B''Au;$$

in which case $\|B''A\|\leq\|B''\|\,\|A\|<\varepsilon$ and the sum is a degenerate operator. By definition the operator BA is completely continuous. Further,

$$ABu=\sum_{k=1}^{n}(u,\ \psi_k)A\varphi_k+AB''u;\quad \|AB''\|\leq\|A\|\,\|B''\|<\varepsilon,$$

and operator AB is also completely continuous.

The Fredholm integral operator

$$Ku=\int_{\Omega}K(P,\ Q)u(Q)\mathrm{d}\Omega_Q, \tag{3}$$

whose kernel $K(P,\ Q)$ is subject to the condition

$$B^2=\int_{\Omega}\int_{\Omega}|K^2(P,\ Q)|\,\mathrm{d}\Omega_P\mathrm{d}\Omega_Q<\infty. \tag{4}$$

is an important example of an operator which is completely continuous in the space $L_2(\Omega)$. The proof of the complete continuity of the Fredholm operator can be found for example in the author's book [7].

An important property of a completely continuous operator, which is usually given as its definition, is given by the following theorem.

THEOREM 1. *In order that an operator T be completely continuous it is necessary and sufficient that it transform any set with bounded norm into a set which is compact in the sense of the convergence in the given space.*

(The proof of theorem 1 is given, for example, in the author's book: "Direct methods in mathematical physics", Gostekhizdat, 1950.)

THEOREM 2. *If A is a symmetric, positive-bounded-below operator with a discrete spectrum, then its inverse operator $G=A^{-1}$ is completely continuous.*

Let $\varphi_1, \varphi_2, \ldots, \varphi_n, \ldots$ be the orthonormal eigenelements of operator A and let $\lambda_1, \lambda_2, \ldots \lambda_n, \ldots$ be the eigenvalues corresponding to them. We recall (cf. § 31) that $\lambda_n \to 0$ if $n \to \infty$. Let $Au=f$, then $u=Gf$. By formula (6) § 30

$$Gf = \sum_{k=1}^{\infty} \frac{(f, \varphi_k)}{\lambda_k} \varphi_k$$

We put

$$G'f = \sum_{k=1}^{n} \frac{(f, \varphi_k)}{\lambda_k} \varphi_k, \quad G''f = \sum_{k=n+1}^{\infty} \frac{(f, \varphi_k)}{\lambda_k} \varphi_k, \tag{5}$$

so that $G=G'+G''$. Operator G' is degenerate and it suffices for us to prove that the norm of operator G'' can be made as small as required.

Since the elements φ_k are orthonormal

$$\|G''f\|^2 = \sum_{k=n+1}^{\infty} \frac{|(f, \varphi_k)|^2}{\lambda_k^2}. \tag{6}$$

As usual, we consider that the eigenvalues are distributed in order of size. We replace the λ_k in series (12) by the lowest of them λ_{n+1}, which leads to the inequality

$$\|G''f\|^2 \leq \frac{1}{\lambda_{n+1}^2} \sum_{k=n+1}^{\infty} |(f, \varphi_k)|^2,$$

or, using Bessel's inequality,

$$\|G''f\| \leq \frac{1}{\lambda_{n+1}} \|f\|.$$

In this relation the equality sign occurs for $f=\varphi_{n+1}$ and the constant $\frac{1}{\lambda_{n+1}}$ cannot be replaced by a lower one. By definition the norm of the operator $\|G''\| = \frac{1}{\lambda_{n+1}}$, which can be as small as required for sufficiently large n.

§ 77. Equations involving a completely continuous operator

We shall consider the equation

$$u - \lambda T u = f, \tag{1}$$

where u is the required element and f is the given element of some Hilbert space H, T is a completely continuous operator in this space and λ is a numerical parameter. In order to solve this equation we proceed in the following fashion. We shall consider that λ can assume any fixed value with modulus not exceeding some constant R, so that $|\lambda| \leq R$. By formula (2) § 76 we put $T = T' + T''$, where operator T' is degenerate and we require that

$$\|T''\| \leq \frac{1}{2R}.$$

Then

$$\|u - \lambda T'' u\| \geq \|u\| - |\lambda| \|T'' u\| \geq \frac{1}{2} \|u\|; \tag{2}$$

it follows from theorem 5 § 45 that *the operator* $(E-\lambda T'')^{-1}$ *exists and is bounded*. Its form is readily established, for direct verification shows that

$$\sum_{k=0}^{\infty}\lambda^k(T'')^k(E-\lambda T'')u=(E-\lambda T'')\sum_{k=0}^{\infty}\lambda^k(T'')^k u=u$$

and, consequently,

$$(E-\lambda T'')^{-1}u=\sum_{k=0}^{\infty}\lambda^k(T'')^k u. \qquad (*)$$

Series (*) converges for any element u. In fact, by formula (14) § 45

$$\|\lambda^n(T'')^n u\|\leq|\lambda|^n\,\|(T'')^n\|\,\|u\|\leq|\lambda|^n\,\|T''\|^n\,\|u\|.$$

But $|\lambda|\leq R$ and $\|T''\|\leq\frac{1}{2R}$, and so

$$\|\lambda^n(T'')^n u\|\leq\frac{1}{2^n}\,\|u\|,$$

and the convergence of the series follows from the fact that its general term does not exceed the general term of a convergent geometrical series. We note that the operator $(E-\lambda T'')^{-1}$ is defined for the entire space, inasmuch as series (*) converges for any u. Finally

$$\|(E-\lambda T'')^{-1}u\|=\|\sum_{n=0}^{\infty}\lambda^n(T'')^n u\|\leq\sum_{n=0}^{\infty}\|\lambda^n(T'')^n u\|$$

$$\leq\|u\|\sum_{n=0}^{\infty}\frac{1}{2^n}=2\|u\|.$$

The last inequality means that $\|(E-\lambda T'')^{-1}\|\leq 2$; this estimate can also be obtained from inequality (2) and theorem 5 § 45. We rewrite equation (1) in the form

$$u-\lambda T''u-\lambda T'u=f.$$

By applying the operator $(E-\lambda T'')^{-1}$ to both of its sides we transform this equation to the equivalent equation

$$u-\lambda(E-\lambda T'')^{-1}T'u=F_\lambda;\quad F_\lambda=(E-\lambda T'')^{-1}f. \qquad (3)$$

Further, let

$$T'u=\sum_{k=1}^{n}(u,\ \psi_k)\varphi_k.$$

The set of elements φ_k and also the set of elements ψ_k can be considered as linearly independent. Now

$$(1-\lambda T'')^{-1}T'u=\sum_{k=1}^{n}(u,\ \psi_k)\omega_{\lambda,\,k};\quad \omega_{\lambda,\,k}=(E-\lambda T'')^{-1}\varphi_k. \tag{4}$$

The elements $\omega_{\lambda,\,k}$ are linearly independent. In fact, let the constants α_k be such that $\sum_{k=1}^{n}\alpha_k\omega_{\lambda,\,k}=0$. Then

$$(E-\lambda T'')\sum_{k=1}^{n}\alpha_k\omega_{\lambda,\,k}=\sum_{k=1}^{n}\alpha_k(E-\lambda T'')\omega_{\lambda,\,k}=0,$$

or

$$\sum_{k=1}^{n}\alpha_k\varphi_k=0.$$

Since the φ_k are linearly independent, $\alpha_k=0$, which was to be proved.

We shall use the notation

$$C_k=(u,\ \psi_k), \tag{5}$$

Equation (3) then assumes the form

$$u-\lambda\sum_{k=1}^{n}C_k\omega_{\lambda,\,k}=F_\lambda, \tag{6}$$

and we are now concerned with determining the constants C_k.

Using formula (4) we reduce equation (3) to the form

$$u-\lambda\sum_{k=1}^{n}(u,\ \psi_k)\omega_{\lambda,\,k}=F_\lambda.$$

Taking the value of u given by equation (6) and using the linear independence of the $\omega_{\lambda,\,k}$, we obtain a system of linear algebraic equations in the unknowns C_k

$$C_m-\lambda\sum_{k=1}^{n}a_{mk}(\lambda)C_k=b_m(\lambda);\qquad m=1,\ 2,\ \ldots,\ a, \tag{7}$$

which is equivalent to equation (1). Here we have put

$$a_{mk}(\lambda)=(\omega_{\lambda,k},\ \psi_m);\quad b_m(\lambda)=(F_\lambda,\ \psi_m). \tag{8}$$

Obviously the coefficients a_{mk} and, consequently, the determinant

$$D_R(\lambda)=\begin{vmatrix} 1-\lambda a_{11}, & -\lambda a_{12}, & \dots, & -\lambda a_{1n} \\ -\lambda a_{21}, & 1-\lambda a_{22}, & \dots, & -\lambda a_{2n} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ -\lambda a_{n1}, & -\lambda a_{n2}, & \dots, & 1-\lambda a_{nn} \end{vmatrix} \tag{9}$$

are functions of λ which are holomorphic in the circle $|\lambda|\leq R$ of the complex λ plane (in which case $D_R(\lambda)\not\equiv 0$ since, obviously $D_R(0)=1$). Hence it follows that in the circle $|\lambda|\leq R$ the determinant $D_R(\lambda)$ can have only a finite number of roots.

If $D_R(\lambda)=0$, then the homogeneous system obtained from (7) by replacing the right-hand sides by zeros has a non-trivial solution, i.e., a solution for which not all the C_m equal zero. Then, obviously, the homogeneous equation

$$u-\lambda Tu=0$$

has a non-trivial solution and the λ considered is a quantity which is the reciprocal of the eigenvalue of operator T. Such λ are also known as characteristic values. If $D_R(\lambda)\neq 0$, then system (7) and also equation (1) have a unique solution, so that the operator $(E-\lambda T)^{-1}$ exists. We shall denote it by Γ_λ. We shall call those values of λ for which Γ_λ exists *regular values*.

For equation (1) containing a completely continuous operator the so called *Fredholm alternative* holds true: *either* the non-homogeneous equation is soluble and uniquely so for any independent term, f, and then the corresponding homogeneous equation has only the trivial solution *or* the non-homogeneous equation is not soluble for any independent term, and then corresponding homogeneous equation has non-trivial solutions. The first part of the alternative occurs if λ is a regular value and the second if λ is a characteristic value.

As we shall see only a finite number of characteristic values of equation (1) are contained in a circle of any radius $|\lambda| \leq R$; hence it follows that there is at the most a countable set of these values in the entire complex λ-plane. All the remaining values of λ are regular.

If the operator Γ_λ exists, then it is bounded. We now prove this. By solving system (7) and substituting its solution in (6) we find after simple manipulation

$$u = \Gamma_\lambda f = (E - \lambda T'')^{-1} f + \frac{1}{D_R(\lambda)} \sum_{k,\, m=1}^{n} \Delta_{km}(\lambda) ((E - \lambda T'')^{-1} f,\ \psi_k) \omega_{\lambda,\, m}, \tag{10}$$

where $\Delta_{km}(\lambda)$ is the cofactor of the element in the kth row and the mth column of the determinant $D_R(\lambda)$.

We saw above that $\|(E - \lambda T'')^{-1}\| \leq 2$. Now formula (10) gives

$$\|\Gamma_\lambda f\| \leq 2\|f\| + \sum_{k,\, m=1}^{n} \left| \frac{\Delta_{km}(\lambda)}{D_R(\lambda)} \right| |((E - \lambda T'')^{-1} f,\ \psi_k)|\ \|\omega_{\lambda,\, m}\|$$

and by Cauchy's inequality

$$\begin{aligned} |((E - \lambda T'')^{-1} f,\ \psi_k)| &\leq \|(E - \lambda T'')^{-1} f\|\ \|\psi_k\| \\ &\leq \|(E - \lambda T'')^{-1}\|\ \|f\|\ \|\psi_k\| \leq 2\|\psi_k\|\ \|f\| \end{aligned}$$

and so

$$\|\Gamma_\lambda f\| \leq 2 \left\{ 1 + \sum_{k,\, m=1}^{n} \left| \frac{\Delta_{km}(\lambda)}{D_R(\lambda)} \right| \|\omega_{\lambda,\, m}\|\ \|\psi_k\| \right\} \|f\|.$$

Thus

$$\|\Gamma_\lambda\| \leq 2 \left\{ 1 + \sum_{k,\, m=1}^{n} \left| \frac{\Delta_{km}(\lambda)}{D_R(\lambda)} \right| \|\omega_{\lambda,m}\|\ \|\psi_k\| \right\}.$$

THEOREM 1. *Let $\{T_n\}$ be a set of completely continuous operators in some Hilbert space H which tend to some completely continuous operator T in the sense that*

$$\|T - T_n\| \underset{n \to \infty}{\to} 0. \tag{11}$$

Further, let f_n be a set of elements of the same space which tend to some element f. If λ is a regular value of the equation

$$u-\lambda Tu=f, \tag{12}$$

then, for sufficiently large n, λ will also be a regular value for the equation

$$u_n-\lambda T_n u_n=f_n, \tag{13}$$

and the solution of equation (13) *will tend to the solution of equation* (12) *as $n\to\infty$.*

We consider the equation

$$v-\lambda T_n v=g, \tag{14}$$

where g is an arbitrary element from H. We write equation (14) in the form

$$v-\lambda Tv-\lambda(T_n-T)v=g. \tag{15}$$

From the conditions of the theorem, we know that the operator Γ_λ exists and applying it to both sides of equation (15), we obtain

$$v-\lambda\Gamma_\lambda(T_n-T)v=\Gamma_\lambda g. \tag{16}$$

Further, we have

$$\|\lambda\Gamma_\lambda(T_n-T)\|\leq|\lambda|\,\|\Gamma_\lambda\|\,\|T_n-T\|, \tag{17}$$

which from the conditions of the theorem can be as small as required for sufficiently large n. Let n be so large that

$$\|\lambda\Gamma_\lambda(T_n-T)\|<\frac{1}{2}.$$

By applying the argument given above to series (*), we can satisfy ourselves that the operator $[E-\lambda\Gamma_\lambda(T_n-T)]^{-1}$ exists, is defined for the entire space and that its norm does not exceed 2.

Now from (16) we find the solution of equation (14):

$$v=[E-\lambda\Gamma_\lambda(T_n-T)]^{-1}\Gamma_\lambda g.$$

Hence it is clear that for the given value of n the operator

$$\Gamma_{n\lambda}=(E-\lambda T_n)^{-1}=[E-\lambda\Gamma_\lambda(T_n-T)]^{-1}\Gamma_\lambda \tag{18}$$

exists; thus we have proved that the value of λ considered is regular for equation (13).

In order to establish the second part of the theorem we shall prove, first of all, that $\|\Gamma_{n\lambda}-\Gamma_\lambda\| \xrightarrow[n\to\infty]{} 0$. For brevity we use the notation

$$\lambda\Gamma_\lambda(T_n-T)=B_n;$$

It is clear from (17) that $\|B_n\|\to 0$ as $n\to\infty$. Now by formulae (18) and by (*)

$$\Gamma_{n\lambda}u=\sum_{k=0}^{\infty}B_n^k\Gamma_\lambda u,$$

whence

$$\Gamma_{n\lambda}u-\Gamma_\lambda u=\sum_{k=1}^{\infty}B_n^k\Gamma_\lambda u$$

and

$$\|\Gamma_{n\lambda}-\Gamma_\lambda\|\leq\sum_{k=1}^{\infty}\|B_n\|^k\|\Gamma_\lambda\|=\frac{\|B_n\|\,\|\Gamma_\lambda\|}{1-\|B_n\|}\xrightarrow[n\to\infty]{}0. \qquad (19)$$

Now the second part of the theorem is readily proved. We have

$$u_n-u=\Gamma_{n\lambda}f_n-\Gamma_\lambda f=(\Gamma_{n\lambda}-\Gamma_\lambda)f_n+\Gamma_\lambda(f_n-f),$$

hence

$$\|u_n-u\|\leq\|\Gamma_{n\lambda}-\Gamma_\lambda\|\,\|f_n\|+\|\Gamma_\lambda\|\,\|f_n-f\|.$$

Since $\|f_n\|\to\|f\|$, $\|f_n\|$ is bounded; by virtue of (19) the first term tends to zero. That the second term tends to zero is obvious. Finally $\|u_n-u\| \xrightarrow[n\to\infty]{} 0$ and the theorem is proved.

THEOREM 2. *If $\|T_n-T\|\to 0$ where T and T_n are completely continuous operators, then the characteristic values of the equation $u-\lambda Tu=0$ are obtained by the limit process as $n\to\infty$ from the characteristic values of the equation $u_n-\lambda T_n u_n=0$.*

The proof is founded on the simple observation that for a sufficiently small change of the operator T the elements of the determinant $D_R(\lambda)$ and, consequently, the determinant itself change by as little as required. Let λ_0 be any root of $D_R(\lambda)$ which lies within the circle $|\lambda|\leq R$ and let ϱ be its

multiplicity. We surround λ_0 by a circle of sufficiently small radius ϱ. Then there will be no roots of $D_R(\lambda)$ besides λ_0 within and on a circle with this radius. In particular on the circle $|\lambda-\lambda_0|=\varrho$ the determinant $D_R(\lambda)$ is non-zero. We shall use the notation

$$q=\min_{|\lambda-\lambda_0|=\varrho}|D_R(\lambda)|; \qquad q>0.$$

We now select an N so great that for $n\geq N$ on the circle $|\lambda-\lambda_0|=\varrho$

$$|D_R(\lambda)-D_R^{(n)}(\lambda)|<q,$$

where $D_R^{(n)}(\lambda)$ is the determinant (9) corresponding to the operator T_n. By Rouche's theorem (cf. e. g. V. I. Smirnov [1], p. 89 or B. A. Fuchs and B. V. Shabat [1], p. 303) $D_R^{(n)}(\lambda)$ has p equal roots in the circle $|\lambda-\lambda_0|<\varrho$. We shall denote them by $\lambda_{n1}, \lambda_{n2}, \ldots, \lambda_{np}$. Obviously

$$|\lambda_{nj}-\lambda_0|<\varrho, \qquad j=1, 2, \ldots, p, \; n\geq N.$$

Since ϱ can be chosen as small as required, the last inequality is equivalent to the assertion that

$$\lim_{n\to\infty}\lambda_{nj}=\lambda_0, \qquad j=1, 2, \ldots, p.$$

§ 78. A sufficient criterion for the convergence of the Bubnov–Galerkin method

This has been discussed by the author ([4], [8]).

Here we shall discuss the convergence of the Bubnov–Galerkin method upon the assumption that the operator A in the equation $Au=f$ has the form

$$A=A_0+K, \tag{1}$$

where A_0 is an operator which is symmetric and positive-bounded-below in some Hilbert space H. We shall assume the field of definition of operator K is wider than the field of definition of operator A_0 (in symbols $D_K\supset D_{A_0}$), so that the expres-

sion Ku is meaningful whenever the expression A_0u is meaningful. Further restrictions will be imposed upon operator K later.

As in § 46, we introduce the space H_0 in which the scalar product equals

$$[u, v]=(A_0u, v). \tag{2}$$

Following the Bubnov–Galerkin method we shall select a sequence of coordinate elements $\{\varphi_n\}$. *We stipulate that this sequence is within the field of definition of the operator* A_0 *and is complete in* H_0.

Equation (5) § 73 in our case has the form

$$\sum_{k=1}^{n}\{(A_0\varphi_k, \varphi_m)+(K\varphi_k, \varphi_m)\}a_k=(f, \varphi_m), \quad m=1, 2, \ldots, n. \tag{3}$$

We orthonormalize $\{\varphi_n\}$ in the space H_0. As we have frequently observed, this does not affect the approximate solution u_n. System (3) is now somewhat simplified. Firstly

$$(A_0\varphi_k, \varphi_m)=\begin{cases} 0, & k\neq m, \\ 1, & k=m. \end{cases}$$

Further

$$(K\varphi_k, \varphi_m)=(A_0GK\varphi_k, \varphi_m)=[T\varphi_k, \varphi_m],$$

where $G=A_0^{-1}$ and $T=GK$. Finally

$$(f, \varphi_m)=(A_0Gf, \varphi_m)=[f', \varphi_m], \qquad f'=Gf.$$

For brevity we introduce the notation

$$\gamma_{mk}=[T\varphi_k, \varphi_m], \qquad b_m=[f', \varphi_m],$$

and we reduce equation (3) to the form

$$a_m+\sum_{k=1}^{n}\gamma_{mk}a_k=b_m, \qquad m=1, 2, \ldots, n. \tag{4}$$

We now make an assumption which will be important in what follows; we suppose that *the operator* $T=GK$ *is completely continuous in the space* H_0. Being completely continuous, this operator is bounded (cf. § 76) in the space H_0 and by

theorem (3) § 45 can be extended to all this space. We shall consider that such an extension has already been accomplished.

The equation $Au=f$ can be represented in the form

$$A_0u+Ku=f. \tag{5}$$

Applying the operator $A_0^{-1}=G$ to both sides of equation (5) we obtain the new equation

$$u+Tu=f'; \quad f'=Gf. \tag{6}$$

Obviously any solution of equation (5) also satisfies equation (6). The converse may be invalid: it may happen that there exists an element of the space H_0 which satisfies equation (6) but does not belong to D_{A_0}. It is then impossible to insert this element in (5). However, we agree to consider such solutions of equation (6) as generalized solutions of equation (5). With such a convention equations (5) and (6) are equivalent.

THEOREM 1: *The approximate solutions of equation* (5) *constructed by the Bubnov–Galerkin method converge in the energy of operator* A_0 *(in other words in the sense of convergence in the space* H_0*) to the exact solution of this equation if the following conditions are satisfied:*

I. *Equation* (5) *has not more than one solution in* H_0;

II. *The operator* $T=A_0^{-1}K$ *is completely continuous in* H_0.

As mentioned above we assume that the operator A_0 is symmetric and positive-bounded-below and that the system of coordinate functions is complete in H_0.

Proof: As was stated above, equation (5) is equivalent to equation (6) which contains the completely continuous operator T. The Fredholm alternative holds for this equation and from condition I it follows that equation (6) has a solution, obviously belonging to the subspace H_0. By our assumptions, the system of coordinate elements $\{\varphi_n\}$ is complete in H_0; we are thus also entitled to consider it to be orthonormalized in H_0. We expand T_n and f' into orthogonal series,

$$Tu=\sum_{k=1}^{\infty}[Tu, \varphi_k]\varphi_k; \quad f'=\sum_{k=1}^{\infty}[f', \varphi_k]\varphi_k$$

and put

$$T_n u = \sum_{k=1}^{n} [Tu, \varphi_k]\varphi_k; \quad f_n' = \sum_{k=1}^{n} [f', \varphi_k]\varphi_k.$$

It is obvious that $|f' - f_n'| \xrightarrow[n\to\infty]{} 0$. We shall further prove that $|T - T_n| \xrightarrow[n\to\infty]{} 0$.

The operator T is completely continuous in H_0 and can thus be expanded into a sum $T = T' + T''$, where T' is a degenerate operator and $|T''| < \frac{\varepsilon}{2}$, where ε is an arbitrarily small positive number. We have

$$(T - T_n)u = Tu - T_n u = \sum_{k=n+1}^{\infty} [Tu, \varphi_k]\varphi_k$$

$$= \sum_{k=n+1}^{\infty} [T'u, \varphi_k]\varphi_k + \sum_{k=n+1}^{\infty} [T''u, \varphi_k]\varphi_k. \tag{7}$$

The magnitude of the second sum is readily obtained, being given by

$$\left| \sum_{k=n+1}^{\infty} [T''u, \varphi_k]\varphi_k \right|^2 = \sum_{k=n+1}^{\infty} |[T''u, \varphi_k]|^2,$$

which, by Bessel's inequality does not exceed

$$|T''u|^2 \leq |T''|^2 \, |u|^2 < \frac{\varepsilon^2}{4} |u|^2.$$

Hence

$$\left| \sum_{k=n+1}^{\infty} [T''u, \varphi_k]\varphi_k \right| < \frac{\varepsilon}{2} |u|.$$

We now consider the first sum in (7). The degenerate operator $T'u$ has the form

$$T'u = \sum_{j=1}^{s} [u, \psi_j]\omega_j,$$

where s is a finite number and ψ_j and ω_j are elements of the space H_0.

We now have

$$\left\|\sum_{k=n+1}^{\infty}[T'u,\ \varphi_k]\varphi_k\right\|=\left\|\sum_{k=n+1}^{\infty}\sum_{j=1}^{s}[u,\ \psi_j][\omega_j,\ \varphi_k]\varphi_k\right\|$$

$$=\left\|\sum_{j=1}^{s}[u,\ \psi_j]\sum_{k=n+1}^{\infty}[\omega_j,\ \varphi_k]\varphi_k\right\|\leq\sum_{j=1}^{s}|[u,\ \psi_j]|\left\|\sum_{k=n+1}^{\infty}[\omega_j,\ \varphi_k]\varphi_k\right\|$$

$$\leq\|u\|\sum_{j=1}^{s}\|\psi_j\|\sqrt{\left(\sum_{k=n+1}^{\infty}|[\omega_j,\ \varphi_k]|^2\right)}.$$

The series $\sum_{k=1}^{\infty}|[\omega_j,\ \varphi_k]|^2$ converges. Hence it is readily concluded that the coefficient of $\|u\|$ will be less than $\frac{\varepsilon}{2}$ for n sufficiently great, say for $n\geq N$. Hence from relations (7) and (8) it also follows that for $n\geq N$, $\|(T-T_n)u\|<\varepsilon\|u\|$ or $\|T-T_n\|<\varepsilon$ which was to be proved.

The operator T_n is degenerate. In fact, it is easily seen that the functional $[T_u,\ \varphi_k]$ is bounded in H_0. By theorem 1 § 45, there exists an element $\tilde{\varphi}_k\in H_0$ such that $[Tu,\ \varphi_k]=[u,\ \tilde{\varphi}_k]$ and

$$T_nu=\sum_{k=1}^{n}[u,\ \tilde{\varphi}_k]\varphi_k. \tag{8}$$

Being degenerate, the operator T_n is completely continuous. Now from theorem 1 § 77 it follows that the equation

$$u_n+T_nu_n=f_n' \tag{9}$$

has a unique solution for sufficiently great n and that

$$u_n\xrightarrow{\mathrm{E}}u, \tag{10}$$

where u is the solution of equation (6).

Substituting for T_n and f_n in (9), we obtain

$$u_n=\sum_{k=1}^{n}\{[f',\ \varphi_k]+[Tu_n,\ \varphi_k]\}\varphi_k=\sum_{k=1}^{n}A_k\varphi_k, \tag{11}$$

where

$$A_k=[f',\ \varphi_k]+[Tu_n,\ \varphi_k].$$

Replacing u_n here by its value as given by formula (11), we obtain a system of equations for the coefficients A_k:

$$A_m + \sum_{k=1}^{n} [T\varphi_k, \varphi_m] A_k = [f', \varphi_m], \qquad m = 1, 2, \ldots, n,$$

which is identical to system (4). Hence it immediately follows that u_n is the Bubnov–Galerkin approximate solution of equation (5); the proposition of theorem 1 now follows from relation (10).

THEOREM 2. *If the operator K is bounded in H and $G = A_0^{-1}$ is completely continuous in H, then the operator $T = GK$ is completely continuous in H_0.*

Proof. The operator is completely continuous in H as it is the product of a bounded and a completely continuous operator [cf. subsection 2, § 77]. Now consider a set M which is bounded in H_0 defined by: $|u| < C$, if $u \in M$. By inequality (7) § 6, $\|u\| < \dfrac{C}{\gamma}$. Since T is completely continuous in H, it is possible to select from M a sequence $\{u_n\}$ such that $\|Tu_m - Tu_n\|_{m, n \to \infty} \to 0$.

We now estimate $|Tu_n - Tu_m|$. By formula (4) § 46

$$|Tu_n - Tu_m|^2 = (A_0 T(u_n - u_m),\ T(u_n - u_m)) = (K(u_n - u_m),\ T(u_n - u_m)).$$

By Cauchy's inequality

$$|Tu_n - Tu_m|^2 \leq \|K(u_n - u_m)\|\, \|T(u_n - u_m)\| = \|K(u_n - u_m)\|\, \|Tu_n - Tu_m\|.$$

Furthermore,

$$\|K(u_n - u_m)\| \leq \|K\|\, \|u_n - u_m\| \leq \|K\|(\|u_n\| + \|u_m\|) < \frac{2C\|K\|}{\gamma}$$

and finally

$$|Tu_n - Tu_m|^2 < \frac{2C\|K\|}{\gamma} \|Tu_n - Tu_m\| \xrightarrow[n, m \to \infty]{} 0.$$

By theorem 1 § 76 the operator T is completely continuous in H_0. By putting $K = E$ (E is the identity operator), we obtain

a corollary. *If the operator* $G=A_0^{-1}$ *is completely continuous in* H, *then it is completely continuous in* H_0.

From theorem 2 there follows a considerably narrower but a simpler criterion for the convergence of the Bubnov–Galerkin method which we formulate thus:

THEOREM 3: The *Bubnov–Galerkin method converges if:*

I. *The equation* $Au=f$ *has no more than one solution*

II. *The operator* $G=A_0^{-1}$ *is completely continuous and the operator* K *is bounded in* H.

NOTE: From theorem 1, there immediately follows the applicability of the Bubnov–Galerkin method to equations of the form $u-\lambda Tu=f$, where T is a completely continuous operator in H, in particular, to a Fredholm integral equation.

Let condition II of theorem I be satisfied. We consider the homogeneous equation

$$A_0u-\lambda Ku=0 \tag{12}$$

which is equivalent to the equation

$$u-\lambda Tu=0. \tag{13}$$

By repeating the reasoning of theorem 1, we can convince ourselves that the application of the Bubnov–Galerkin method to the problem of finding the eigenvalues of equation (12) is equivalent to finding the eigenvalues of the equation

$$u_n-\lambda T_nu_n=0. \tag{14}$$

By theorem 2 § 78 the eigenvalues of equation (13) are the limits of the corresponding eigenvalues of equation (14). Hence there follows.

THEOREM 4. *If the operator* $T=A_0^{-1}K$ *is completely continuous in* H_0 *then the application of the Bubnov – Galerkin method to the problem of seeking the eigenvalues leads to a convergent process.*

NOTE 1. Let A be a positive-bounded-below, symmetric operator with a discrete spectrum. When applied to the problem of the eigenvalues of the equation

$$Au-\lambda u=0 \tag{15}$$

the Bubnov–Galerkin and Ritz methods give identical results. But then by theorem (4) if $\lambda_k^{(n)}$ is the Ritz approximation to the kth eigenvalue λ_k of operator A, then $\lambda_k^{(n)} \xrightarrow[n\to\infty]{} \lambda_k$. We recall that in § 22 this equality was proved only for the first eigenvalue λ_1.

NOTE 2. In the works of G.I. Petrov[1] and M. V. Keldysh[1], it is asserted that the eigenelements of equation (12) can be constructed as the limits of the "approximate" eigenelements, obtained by the Bubnov–Galerkin method. This assertion, as was shown by N. I. Pol'sky [1], is invalid. Under certain conditions, it can happen that it is impossible to approximate to the eigenelements of equation (12) by the Galerkin method.

In order to convince ourselves of this we consider an example which we borrow from Pol'sky's paper (*loc. cit.*). We consider the space $L_2(0, 1)$, we put $A_0 = E$ in (12) and

$$Ku = \int_0^1 K(x, t)u(t)\mathrm{d}t = \int_0^1 \{a_1(x)b_1(t) + a_2(x)b_2(t)\}u(t)\mathrm{d}t,$$

where $0 < q < 1$;

$$a_1(x) = \sin \pi x,\ b_1(x) = 2 \sin \pi x + \frac{q^4 \sin 3\pi x}{\sqrt{(1-q^2)}} + \sum_{k=4}^{\infty} q^k \sin k\pi x,$$

$$a_2(x) = 2 \sin 2\pi x + \frac{q^4 \sin 3\pi x}{\sqrt{(1-q^2)}} - \sum_{k=4}^{\infty} q^k \sin k\pi x,\ b_2(x) = \sin 2\pi x.$$

The conditions of theorem 4 are obviously satisfied and $H_0 \equiv M$.

Using the readily verified relations

$$(a_1, b_1) = (a_2, b_2) = 1,\ (a_1, b_2) = (a_2, b_1) = 0,$$

we find that the equation

$$u(x) - \lambda \int_0^1 K(x, t)u(t)\mathrm{d}t = 0 \qquad (*)$$

has a unique eigenvalue $\lambda = 1$ to which there correspond two linearly independent eigenfunctions $a_1(x)$ and $a_2(x)$. We shall

now solve equation (*) by the Bubnov–Galerkin method, putting $u_n = \sum_{k=1}^{n} C_k \varphi_k(x)$; $\varphi_k(x) = \sqrt{2} \sin k\pi x$. Let $\lambda^{(n)}$ be the corresponding approximate eigenvalue obtained by the Bubnov–Galerkin method. We may readily satisfy ourselves that u_n and $\lambda^{(n)}$ satisfy the equation

$$u_n(x) - \lambda^{(n)} \int_0^1 [a_1^{(n)}(x) b_1(t) + a_2^{(n)}(x) b_2(t)] u_n(t) \mathrm{d}t = 0, \quad (**)$$

where

$$a_1^{(n)}(x) = \sum_{k=1}^{n} \varphi_k(x)(a_1, \varphi_k) \xrightarrow[n\to\infty]{} a_1(x),$$

$$a_2^{(n)}(x) = \sum_{k=1}^{n} \varphi_k(x)(a_2, \varphi_k) \xrightarrow[n\to\infty]{} a_2(x).$$

It is readily verified that equation (**) has the unique eigenvalue $\lambda^{(n)} = 1$ to which there corresponds only one eigenfunction $a_1^{(n)}(x) = a_1(x)$. Thus the Bubnov–Galerkin method does not permit us to approximate to the eigenfunction $a_2(x)$ of equation (*) in our example.

As N. I. Pol'sky has shown, all the eigenelements of equation (12) can be obtained as the limits of the "approximate" eigenelements if the resolvent of the operator $A_0^{-1}K$ has only simple poles. This occurs, in particular, if $K = E$, i.e. under the conditions of the Ritz method.

§ 79. Application to ordinary differential equations

In § 75 we established the convergence of the Bubnov–Galerkin method for ordinary second-order differential equations. The general theorems of § 78 readily enable us to discuss a considerably more general class of equations. As an example we consider the equation

$$(-1)^m u^{(2m)} - \lambda K u = f(x), \quad (1)$$

where Ku is a linear differential operator of order $2m-1$. We shall seek the solution of this equation which, at the ends of the segment $a \leq x \leq b$ satisfies the boundary conditions

$$u(a) = u'(a) = \ldots = u^{(m-1)}(a) = 0$$
$$u(b) = u'(b) = \ldots = u^{(m-1)}(b) = 0. \tag{2}$$

We shall assume that the problem posed has a unique solution. We shall prove that the Bubnov – Galerkin method for this problem leads to a convergent process.

We put $A_0 u = (-1)^m u^{(2m)}$. As was proved in § 20, the operator A_0 is positive-bounded-below for the set of functions which are differentiable a sufficient number of times and which satisfy boundary conditions (2). As is known the operator $G = A_0^{-1}$ has the form

$$Gf = \int_a^b G(x, \xi) f(\xi) \mathrm{d}\xi,$$

where $G(x, \xi)$ is the Green function for the equation $u^{(2m)} = 0$ which satisfies boundary conditions (2). This function is continuous together with its derivatives up to order $2m-2$ inclusive in the square $a \leq x, \xi \leq b$ and its derivatives of order $2m-1$ are discontinuous only on the diagonal $x = \xi$, where they undergo finite jumps. We introduce the space H_0 in which this time

$$[u, v] = \int_a^b (-1)^m u^{(2m)} v \mathrm{d}x.$$

Integrating by parts m times and using conditions (2),

$$[u, v] = \int_a^b u^{(m)} v^{(m)} \mathrm{d}x$$

and, consequently,

$$\|u\|^2 = \int_a^b [u^{(m)}]^2 \mathrm{d}x. \tag{3}$$

We now consider the operator Ku. We suppose that the coefficients of the operator K are differentiable a sufficient

number of times. We put $GKu=v(x)$. Obviously v satisfies boundary conditions (2). Further, integrating by parts in order to exclude derivatives of n of order higher than m and again using conditions (2), we reduce the expression for v to the form

$$v(x)=\int_a^b G_1(x,\ \xi)u^{(m)}(\xi)\mathrm{d}\xi$$

$$+\int_a^b G(x,\ \xi)[p_{m-1}(\xi)u^{(m-1)}(\xi)+\ldots+p_0(\xi)u(\xi)]\mathrm{d}\xi, \qquad (4)$$

where the kernel $G_1(x,\ \xi)$ has m bounded derivatives.

Differentiating (4) m times with respect to x, we have

$$v^{(m)}(x)=\int_a^b \frac{\partial^m G_1(x,\ \xi)}{\partial x^m}u^{(m)}(\xi)\mathrm{d}\xi$$

$$+\int_a^b \frac{\partial^m G(x,\xi)}{\partial x^m}[p_{m-1}(\xi)u^{m-1}(\xi)+\ldots+p_0(\xi)u(\xi)]\mathrm{d}\xi. \qquad (5)$$

In integral (5) we express $u,\ u',\ \ldots,\ u^{(m-1)}$ in terms of $u^{(m)}$ according to the well-known formula

$$u^{(i)}(\xi)=\frac{1}{(m-j-1)!}\int_a^\xi (\xi-\eta)^{m-j-1}u^{(m)}(\eta)\mathrm{d}\eta, \qquad (6)$$

which is valid if u satisfies the first m of conditions (2). Substituting the values of $u^{(i)}$ from (6) in equation (5), and changing the order of integration, we find that $v^{(m)}(x)$ has the form

$$v^{(m)}(x)=\int_a^b N(x,\ \xi)u^{(m)}(\xi)\mathrm{d}\xi, \qquad (7)$$

where $N(x,\ \xi)$ is a continuous function. Integral (7) is a Fredholm operator acting upon the function $u^{(m)}$. As was stated in § 76, this is a completely continuous operator upon $u^{(m)}$, con-

sidered as an element of the space $L_2(a, b)$. Now consider a set of functions $u(x)$ which is bounded in H_0. This means that

$$\|u\|^2 = \int_a^b |u^{(m)}(x)|^2 dx < C; \quad C = \text{const.}$$

By theorem 1 § 76 it is possible to select from this set a sequence $\{v_n(x)\}$ such that $\|v_n^{(m)} - v_k^{(m)}\|^2 = \int_b^a |v_n^{(m)}(x) - v_k^{(m)}(x)|^2 dx \xrightarrow[n,k\to\infty]{} 0$. But then it follows from (3) that $\|v_n - v_k\| \xrightarrow[n,k\to\infty]{} 0$. Hence it is clear that the operator on the right-hand side of (4) is completely continuous in H_0.

On the basis of theorem 1 § 78, the Bubnov – Galerkin method gives a sequence which is convergent in H_0. If we denote by u_n the approximate solution obtained by the Bubnov–Galerkin method, and by u the exact solution of the problem, then for $k < m$, $u_n^{(k)}(x) \xrightarrow[n\to\infty]{} u^{(k)}(x)$ uniformly and $u_n^{(m)}(x) \xrightarrow[n\to\infty]{} u^{(m)}(x)$ in the mean.

Theorem 4 § 78 allows us to assert the convergence of the Bubnov–Galerkin method when applied to the problem of the eigenvalues of the homogeneous problem (1).

The results mentioned in this section were obtained in a considerably more complex manner by M.V. Keldysh [1].

§ 80. The Dirichlet problem for a second-order elliptic equation

We shall seek the integral over some finite domain Ω of the equation with (in general) variable coefficients

$$-\sum_{i,k=1}^{m} \frac{\partial}{\partial x_i}\left(A_{ik}\frac{\partial u}{\partial x_k}\right) + \sum_{i=1}^{m} B_i \frac{\partial u}{\partial x_i} + Cu = f(P), \qquad (1)$$

which vanishes on the boundary S of the domain Ω.

We shall consider that equation (1) is elliptic non-degenerate. This implies the existence of a constant μ_0 such that for

any point $P \in \overline{\Omega}$ and for any real numbers $t_1, t_2, \ldots, t_m$ the following inequality holds

$$\sum_{i,k=1}^{m} A_{ik} t_i t_k \geq \mu_0 \sum_{i=1}^{m} t_i^2. \tag{2}$$

We shall further suppose that the coefficients A_{ik} and their first derivatives are continuous in Ω and that the coefficients B_i and C are continuous in $\overline{\Omega}$. (The restrictions imposed upon the coefficients can in fact be relaxed.) We suppose also that the function $f(P)$ has a finite norm. Finally we assume that the problem posed has a unique solution. We put

$$A_0 u = -\sum_{i,k=1}^{m} \frac{\partial}{\partial x_i}\left(A_{ik}\frac{\partial u}{\partial x_k}\right); \quad Ku = \sum_{i=1}^{m} B_i \frac{\partial u}{\partial x_i} + Cu. \tag{3}$$

In § 24, we established that the operator A_0 is positive-bounded-below for the set of functions which vanish on S. For such functions the inequality

$$(A_0 u, u) \geq \gamma^2 \|u\|^2, \quad \gamma > 0 \tag{4}$$

holds. We now introduce the space H_0 by putting

$$[u, v] = (A_0 u, v) = \int_{\Omega} \sum_{i,k=1}^{m} A_{ik} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_k} \, d\Omega.$$

We have explained in § 37 that the operator A_0 has a discrete spectrum. By theorem 2 § 76, the operator $G = A_0^{-1}$ is completely continuous in the space $L_2(\Omega)$ of functions which are quadratically summable in Ω. We shall prove that the operator $T = GK$ is completely continuous in H_0. Consider a set of functions which is bounded in H_0, so that $|u| \leq N = \text{const}$. This means that

$$\int_{\Omega} \sum_{i,k=1}^{m} A_{ik} \frac{\partial u}{\partial x_i} \frac{\partial u}{\partial x_k} \, d\Omega < N^2. \tag{5}$$

From inequalities (2) and (5), it follows that

$$\sum_{i=1}^{m} \left\| \frac{\partial u}{\partial x_i} \right\|^2 = \int_{\Omega} \sum_{i=1}^{m} \left(\frac{\partial u}{\partial x_i} \right)^2 d\Omega < \frac{N^2}{\mu_0}.$$

Moreover, by virtue of inequality (4),

$$\|u\| < \frac{N}{\gamma}.$$

Now

$$\|Ku\| \leq \left\{ \sum_{i=1}^{m} \left\| \frac{\partial u}{\partial x_i} \right\|^2 \right\}^{\frac{1}{2}} \operatorname{Max} \left\{ \sum_{i=1}^{m} B_i^2 \right\} + \|u\| \operatorname{Max} |C(P)| \leq C_1 N ,$$

$$C_1 = \text{const.} \tag{6}$$

Since G is completely continuous in $H = L_2(\Omega)$ and the set of functions Ku, are as we see, bounded in H, it is possible to select a sequence $\{u_n\}$ such that there exists in the space H the limit

$$\lim_{n \to \infty} GKu_n = \lim_{n \to \infty} Tu_n.$$

Hence it follows that

$$\lim_{k,\, n \to \infty} \|Tu_n - Tu_k\| = 0. \tag{7}$$

We now estimate $\|Tu_n - Tu_k\|$. We have, by formula (1) § 8

$$\|Tu_n - Tu_k\|^2 = (A_0 T(u_n - u_k),\ T(u_n - u_k)) = (Ku_n - Ku_k,\ Tu_n - Tu_k),$$

and further, by Cauchy's inequality,

$$\|Tu_n - Tu_k\|^2 \leq \|Ku_n - Ku_k\| \, \|Tu_n - Tu_k\| \leq 2C_1 N \|Tu_n - Tu_k\|.$$

By virtue of inequality (7)

$$\lim_{k,\, n \to \infty} \|Tu_n - Tu_k\| = 0. \tag{8}$$

Thus, if the set of functions is bounded in H_0, it is then possible to separate from it a sequence $\{u_n\}$ such that equation (8) is satisfied. This means that the limit

$$\lim_{n\to\infty} Tu_n$$

(in the space H_0) exists. Thus we have proved that the operator T is completely continuous in H_0 and, consequently, that the Bubnov – Galerkin method converges for our problem. In precisely the same way, we can prove its convergence for the eigenvalue problem.

The problem of this section has been studied by M.V. Keldysh who arrived at the same result by considerably more complex means and under the additional assumption that the domain boundary is sufficiently smooth.

NOTE: We established in § 78 that in the general case $\|u_n-u\|\to 0$ where u_n is the approximate solution obtained by the Bubnov – Galerkin method. In the present case this means that the first derivatives of u_n converge in the mean to the corresponding derivatives of u.

It is also possible under known conditions to justify the Bubnov – Galerkin method for degenerate elliptic equations. This question has been investigated in the author's paper [15].

To illustrate the method we consider the equation

$$\nabla^2 u + x\frac{\partial u}{\partial x} = 2x^2 + 2y^2 + 2x^2y^2 - 2x^2y - xy^2 + xy - 2x - 2y.$$

It has the solution

$$u = xy(x-1)(y-1),$$

which vanishes on the edges of the square

$$x=0, \quad x=1, \quad y=0, \quad y=1. \tag{9}$$

We shall construct an approximation to this solution by the Bubnov – Galerkin method taking $\sin k\pi x \sin l\pi y$; $k, l = 1, 2, 3, \ldots$ as the coordinate functions. These also vanish on the

edges of square (9). We restrict ourselves to terms for which $k+l \leq 4$ and put

$$u_6 = a_1 \sin \pi x \sin \pi y + a_2 \sin 2\pi x \sin \pi y + a_3 \sin \pi x \sin 2\pi y + a_4 \sin 3\pi x \sin \pi y + a_5 \sin 2\pi x \sin 2\pi y + a_6 \sin \pi x \sin 3\pi y.$$

The Bubnov – Galerkin equations in this case are as follows:

$$\left(\frac{\pi^2}{2}+\frac{1}{8}\right) a_1 + \frac{1}{3} a_2 - \frac{3}{16} a_4 = \frac{4}{\pi^6}(9\pi^2 - 8),$$

$$-\frac{1}{3} a_1 + \left(\frac{5}{4}\pi^2 + \frac{1}{8}\right) a_2 + \frac{3}{5} a_4 = -\frac{2}{\pi^4},$$

$$\left(\frac{5}{4}\pi^2 + \frac{1}{8}\right) a_3 + \frac{1}{3} a_5 = 0,$$

$$\frac{3}{16} a_1 - \frac{3}{5} a_2 + \left(\frac{5}{2}\pi^2 + \frac{1}{8}\right) a_4 = \frac{4}{27\pi^6}(49\pi^2 - 8),$$

$$-\frac{1}{3} a_3 + \left(2\pi^2 + \frac{1}{8}\right) a_5 = 0, \quad \left(\frac{5}{2}\pi^2 + \frac{1}{8}\right) a_6 = \frac{4}{27\pi^6}(41\pi^2 - 8).$$

The solution of this system is:

$a_1 = 0{\cdot}066553$; $a_2 = 0{\cdot}000014499$; $a_3 = 0$;
$a_4 = 0{\cdot}0024528$; $a_5 = 0$; $a_6 = 0{\cdot}0024648$.

We know the exact solution of the problem, therefore we can readily estimate the error of the approximate solution. The energy norm of this error is

$$\|u - u_6\| = \left\{ \int_0^1 \int_0^1 \left[\left(\frac{\partial u}{\partial x} - \frac{\partial u_6}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y} - \frac{\partial u_6}{\partial y}\right)^2 \right] \right\}^{\frac{1}{2}} dx\,dy = 0{\cdot}00729.$$

The energy norm of the exact solution is

$$\left\{ \int_0^1 \int_0^1 [(2x-1)^2 (y-y^2)^2 + (x-x^2)^2 (2y-1)^2]^2 \, dx\,dy \right\}^{\frac{1}{2}} = \frac{1}{\sqrt{45}} = 0{\cdot}149.$$

The relative error of the approximate solution u_6 in the energy norm is

$$\frac{0{\cdot}00729}{0{\cdot}149}\approx 4{\cdot}9\%.$$

§ 81. The Neumann problem and the mixed problem for a second-order elliptic equation

The problem of integrating equation (1) § 80 under the boundary conditions

$$\sum_{i,\,k=1}^{m} A_{ik}\frac{\partial u}{\partial x_k}\cos(n, x_i)=0, \tag{1}$$

where n is the external normal to S is called the Neumann problem. We assume that this problem has a unique solution (obviously it is necessary for this that $C(P)\not\equiv 0$). This time we put

$$\left.\begin{aligned} A_0u &= -\sum_{i,\,k=1}^{m}\frac{\partial}{\partial x_i}\left(A_{ik}\frac{\partial u}{\partial x_k}\right)+u,\\ Ku &= \sum_{i=1}^{m} B_i\frac{\partial u}{\partial x_i}+[C(P)-1]u. \end{aligned}\right\} \tag{2}$$

The operator A_0u is positive-bounded-below for the linear set of functions whose first and second derivatives are continuous in $\overline{\Omega}$ and which satisfy condition (1). In fact, integrating by parts and using condition (1), we obtain

$$(A_0u, u)=\int_{\Omega}\sum_{i,\,k=1}^{m} A_{ik}\frac{\partial u}{\partial x_i}\frac{\partial u}{\partial x_k}\,d\Omega+\int_{\Omega}u^2d\Omega\geq\int_{\Omega}u^2d\Omega=\|u\|^2.$$

The remainder of the argument proceeds as in § 80 and leads to the same conclusion about the convergence of the Bubnov –

Galerkin method. That is, it is possible to assert that the approximate solutions of the Neumann problem for the equation

$$-\sum_{i=1}^{m}\frac{\partial}{\partial x_i}\sum_{k=1}^{m}A_{ik}\frac{\partial u}{\partial x_k}+\sum_{i=1}^{m}B_i\frac{\partial u}{\partial x_i}+Cu=f,$$

constructed by the Bubnov–Galerkin method and their first derivatives converge in the mean to the exact solution of this problem and to its corresponding derivatives. Further the eigenvalues of the Neumann problem for this equation can be obtained as the limits of the approximate eigenvalues obtained by the Bubnov–Galerkin method.

We now turn to the mixed boundary value problem for a second-order elliptic equation. We consider equation 1 § 80 under the boundary condition

$$\sum_{i,k=1}^{m}A_{ik}\frac{\partial u}{\partial x_k}\cos(n,\,x_i)+\sigma(P)u=0, \tag{3}$$

where $\sigma(P)$ is a bounded function for a variable point P on the boundary S which we shall consider to be as smooth as is necessary for the proof.

We denote by A_1u the operator

$$-\sum_{i,k=1}^{m}\frac{\partial}{\partial x_i}\left(A_{ik}\frac{\partial u}{\partial x_k}\right).$$

We shall consider its application to the linear set of functions whose second derivatives are continuous in $\overline{\Omega}$ and which satisfy boundary condition (3). We shall prove that, for this linear set, the operator A_1u is bounded below, i.e., that there exists a constant k such that the inequality

$$(A_1u,\,u)\geq k\|u\|^2. \tag{4}$$

holds (the constant k can also be negative). We have

$$(A_1u, u) = -\int_\Omega u \sum_{i,k=1}^{m} \frac{\partial}{\partial x_i}\left(A_{ik}\frac{\partial u}{\partial x_k}\right)d\Omega$$

$$= \int_\Omega \sum_{i,k=1}^{m} A_{ik}\frac{\partial u}{\partial x_i}\frac{\partial u}{\partial x_k}\,d\Omega - \int_S u \sum_{i,k=1}^{m} A_{ik}\frac{\partial u}{\partial x_k}\cos(n, x_i)dS,$$

or, using condition (3)

$$(A_1u, u) = \int_\Omega \sum_{i,k=1}^{m} A_{ik}\frac{\partial u}{\partial x_i}\frac{\partial u}{\partial x_k}\,d\Omega + \int_S \sigma u^2 dS. \tag{5}$$

Let μ_0 have the same value as in § 80. Further let σ_0 denote the upper bound of the function $|\sigma(P)|$.

From (5) it follows that

$$(A_1u, u) \geq \mu_0 \int_\Omega \sum_{i=1}^{m} \left(\frac{\partial u}{\partial x_i}\right)^2 d\Omega - \sigma_0 \int_S u^2 dS.$$

We now construct the functions $a_i(P)$ which are continuous in $\overline{\Omega}$ and coincide with $\cos(n, x_i)$ on S. Assuming that the surface S is sufficiently smooth, we can convince ourselves that the first derivatives of $a_i(P)$ are also continuous in $\overline{\Omega}$. We now have

$$\int_S u^2 dS = \int_S u^2 \sum_{i=1}^{m} \cos^2(n, x_i)dS = \int_S \sum_{i=1}^{m} u^2 a_i \cos(n, x_i)dS$$

$$= \int_\Omega u^2 \sum_{i=1}^{m} \frac{\partial a_i}{\partial x_i}\,d\Omega + 2\int_\Omega \sum_{i=1}^{m} a_i u \frac{\partial u}{\partial x_i}\,d\Omega. \tag{6}$$

The first integral in (6) is less than

$$C' \int_\Omega u^2 d\Omega, \quad C' = \max \left| \sum_{i=1}^{m} \frac{\partial a_i}{\partial x_i} \right|.$$

We estimate the second in the following way. We denote by C_1 the greatest of the maxima of the functions a_i and by ε, an arbitrary positive number. Then

$$\left| 2\int_\Omega \sum_{i=1}^{n} a_i u \frac{\partial u}{\partial x_i}\,d\Omega \right| < 2C_1 \int_\Omega \frac{|u|}{\varepsilon} \sum_{i=1}^{m} \varepsilon \left| \frac{\partial u}{\partial x_i} \right| d\Omega$$

$$\leq \frac{C_1}{\varepsilon^2} \int_\Omega u^2 d\Omega + C_1 m \varepsilon^2 \int_\Omega \sum_{i=1}^{m} \left(\frac{\partial u}{\partial x_i} \right)^2 d\Omega.$$

(Here we have used the inequality

$$\left(\sum_{k=1}^{m} a_k \right)^2 \leq m \sum_{k=1}^{m} a_k^2, \quad a_k \geq 0,$$

which is obtained from Cauchy's inequality (8_2) § 3 for $b_k = 1$.) Hence

$$(A_1 u,\ u) \geq (\mu_0 - C_1 m \sigma_0 \varepsilon^2) \int_\Omega \sum_{i=1}^{m} \left(\frac{\partial u}{\partial x_i} \right)^2 d\Omega + k \int_\Omega u^2 d\Omega,$$

where

$$k = -\sigma_0 \left(C' + \frac{C_1}{\varepsilon^2} \right).$$

We choose ε such that $\mu_0 - C_1 \sigma_0 \varepsilon^2 < 0$. By dropping the non-negative first term on the right in the last inequality, we arrive at inequality (4).

We define the operator A_0 in the following way:

$$A_0 u = A_1 u, \qquad k > 0$$
$$A_0 u = A_1 u + (1-k)u, \qquad k \leq 0$$

where k is the constant in equality (4).

With such a choice the operator $A_0 u$ is obviously positive-bounded-below. The remainder of the argument proceeds as in § 80 and leads to the same conclusions.

In the particular case where $m = 1$ we arrive at the problem of integrating an ordinary linear second-order differential equation under boundary conditions of the form

$$u'(a) + \alpha u(a) = 0, \quad u'(b) + \beta u(b) = 0$$

(where α, β are some constants).

§ 82. Modification of the Bubnov–Galerkin method for the case of natural boundary conditions

In § 78 we formulated a requirement restricting the choice of the coordinate elements. This was that these elements must belong to the field of definition of the operator. It is possible to state a simple sufficient condition which allows us to relax the stated requirement. *Let the operator K be defined for all the elements of the space H_0. Then it is possible to take any system of elements of this space which is complete in H_0 as the coordinate elements.* The Bubnov–Galerkin system of equations must then be written not in the form (3) § 78, which would be devoid of meaning, but in the form

$$\sum_{k=1}^{n}\{[\varphi_k, \varphi_m]+(K\varphi_k, \varphi_m)\}a_k=(f, \varphi_m), \quad m=1, 2, \ldots, n. \quad (1)$$

We may readily convince ourselves of the validity of the proposition stated here by following the course of the proof of theorem 1 § 78.

As an example, we consider the boundary value problem discussed in § 81 for the equation

$$-\sum_{i, k=1}^{m}\frac{\partial}{\partial x_i}\left(A_{ik}\frac{\partial u}{\partial x_k}\right)+\sum_{i=1}^{m}B_i\frac{\partial u}{\partial x_i}+Cu=f(P) \quad (2)$$

under the mixed boundary condition

$$\left[\sum_{i, k=1}^{m}A_{ik}\frac{\partial u}{\partial x_k}\cos(n, x_i)+\sigma u\right]_S=0. \quad (3)$$

We retain the definitions of the operators A_0 and K given in § 81.

To be explicit, let the constant k in inequality (4) § 81 b greater than zero, then

$$A_0u=-\sum_{i, k=1}^{m}\frac{\partial}{\partial x_i}\left(A_{ik}\frac{\partial u}{\partial x_k}\right), \quad Ku=\sum_{i=1}^{m}B_i\frac{\partial u}{\partial x_i}+Cu.$$

For the operator A_0, boundary condition (3) is natural. The space consists of all functions having quadratically summable first generalized derivatives in Ω, and the operator K is defined in an obvious manner for all the functions from H_0. Therefore, in applying the Bubnov–Galerkin method to our problem, we need not subject the coordinate functions to any boundary conditions.

It is readily seen that in the given case

$$[u, v] = \int_{\Omega} \sum_{i,k=1}^{m} A_{ik} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_k} d\Omega + \int_{S} \sigma uv dS,$$

and so equation (1) for the modified Bubnov–Galerkin method has the form

$$\sum_{k=1}^{n} \left\{ \int_{\Omega} \sum_{i,j=1}^{m} \left(A_{ij} \frac{\partial \varphi_k}{\partial x_i} \frac{\partial \varphi_m}{\partial x_j} + \sum_{i=1}^{m} B_i \frac{\partial \varphi_k}{\partial x_i} \varphi_m + C \varphi_k \varphi_m \right) d\Omega \right.$$

$$\left. + \int_{\Omega} \sigma \varphi_k \varphi_m dS \right\} a_k = \int_{\Omega} f \varphi_m d\Omega, \qquad m = 1, 2, \ldots, n.$$

CHAPTER X

THE METHOD OF LEAST SQUARES

THE PRESENT chapter consists principally of results due to the author (cf. [3] and [6]).

§ 83. Basis of the method

Let A be a linear operator which is defined for some linear set D_A which is dense in a given Hilbert space H. We require to solve the equation

$$Au=f, \tag{1}$$

where f is a given element from H. For this purpose we can use *the method of least squares* which consists of the following. We choose a sequence of linearly independent coordinate elements $\{\varphi_n\}$, $\varphi_n \in D_A$ and construct an approximate solution of equation (1) in the form

$$u_n=\sum_{k=1}^{n} a_k\varphi_k \tag{2}$$

The a_k are constants which are determined from the requirement that the quantity

$$\|Au_n-f\|^2 \tag{3}$$

is a minimum.

First of all we must say a word about terminology. We shall call the sequence $\{\varphi_n\}$ $A-$complete if for any $u \in D_A$ and for any $\varepsilon>0$, it is possible to find a positive integer n and constants $\alpha_1, \alpha_2, \ldots, \alpha_n$, such that

$$\|Au-Au_n\|<\varepsilon, \text{ where } u_n=\sum_{k=1}^{n}\alpha_k\varphi_k.$$

We note that A-completeness follows from ordinary completeness if the operator A is bounded. In fact, if the system $\{\varphi_n\}$ is complete and the operator A is bounded, then it is possible to find $n, \alpha_1, \alpha_2, \ldots, \alpha_n$ such that $\|u-u_n\| < \dfrac{\varepsilon}{\|A\|}$. It follows that

$$\|Au - Au_n\| = \|A(u-u_n)\| \leq \|A\| \, \|u-u_n\| < \varepsilon.$$

Condition (3) leads to a system of linear equations in the unknowns $a_1, a_2, \ldots, a_n$. We shall now establish the form of this system. We have

$$\|Au_n - f\|^2 = \left(\sum_{k=1}^{n} a_k A\varphi_k - f, \sum_{m=1}^{n} a_m A\varphi_m - f \right)$$

$$= \sum_{k,\, m=1}^{n} a_k \bar{a}_m (A\varphi_k A\varphi_m) - \sum_{k=1}^{n} a_k (A\varphi_k f) - \sum_{k=1}^{n} \bar{a}_k (f, A\varphi_k) + (f, f). \quad (4)$$

(In general we consider the Hilbert space to be complex.)

We put $a_k = \alpha_k + i\beta_k$. The values of α_k and β_k giving the minimum value of $\|Au_n - f\|^2$ satisfy the equations

$$\frac{\partial \|Au_n - f\|^2}{\partial \alpha_m} = 0, \quad \frac{\partial \|Au_n - f\|^2}{\partial \beta_m} = 0, \quad m = 1, 2, \ldots, n$$

or, what is the same thing, the equations

$$\frac{1}{2} \left\{ \frac{\partial \|Au_n - f\|^2}{\partial \alpha_m} + i \frac{\partial \|Au_n - f\|^2}{\partial \beta_m} \right\} = \frac{\partial \|Au_n - f\|^2}{\partial \bar{a}_m} = 0.$$

Differentiating (4) with respect to $\bar{a}_m$, we obtain the system that we are interested in

$$\sum_{k=1}^{n} a_k (A\varphi_k, A\varphi_m) = (f, A\varphi_m), \quad m = 1, 2, \ldots, n. \quad (5)$$

We note that system (5) is symmetric. The determinant of system (5) is the Gram determinant of the elements $A\varphi_1$, $A\varphi_2, \ldots, A\varphi_n$.

We suppose that equation (1) can have not more than one solution, in other words, that the homogeneous equation

$Au=0$ has only one trivial solution $u=0$. Then $A\varphi_1$, $A\varphi_2$, ..., $A\varphi_n$ are linearly independent. In the opposite case it would be possible to find constants $c_1, c_2, \ldots, c_n$, not all zero, such that

$$c_1A\varphi_1+c_2A\varphi_2+\ldots+c_nA\varphi_n=A\left(\sum_{k=1}^{n}c_k\varphi_k\right)=0.$$

But then

$$\sum_{k=1}^{n}c_k\varphi_k=0,$$

which contradicts the initial assumption about the linear independence of the coordinate elements. By theorem 2 § 11, the determinant of system (5) is non-vanishing and the system mentioned is soluble uniquely.

We now formulate the result obtained in the form of a lemma.

LEMMA 1. *If the homogeneous equation $Au=0$ has only a trivial solution (in other words, if the inverse operator A^{-1} exists) then approximate solutions can be constructed by the method of least squares for any n and they are determined uniquely.*

Sufficient conditions for the convergence of the approximate solutions obtained by the method of least squares are given by the following lemma.

LEMMA 2. *The method of least squares gives a sequence of approximate solutions which converge to the exact solution if the following conditions are satisfied:*

A. The sequence of coordinate elements is A-complete

B. Equation (1) *is soluble*

C. There exists a constant K such that for any $u\in D_A$

$$\|u\|\leq K\|Au\|. \tag{6}$$

Condition C guarantees the existence (and the boundedness) of the inverse operator A^{-1}. By lemma 1, a sequence of approximate solutions $\{u_n\}$ can be constructed. We shall denote by u the solution of equation (1) which exists by virtue of

condition B. Since the sequence $\{\varphi_n\}$ is A-complete (condition A) it is possible, for a given $\varepsilon>0$ to find $n_0, \alpha_1, \alpha_2, \ldots, \alpha_{n_0}$ such that

$$\left\| Au - \sum_{k=1}^{n_0} \alpha_k A\varphi_k \right\| < \frac{\varepsilon}{K}.$$

The last inequality remains valid if $\sum_{k=1}^{n_0} \alpha_k \varphi_k$ is replaced by u_n as determined from (5), since the left-hand side of this inequality is then a minimum.

This minimum does not increase with increasing n_0, so that for $n \geq n_0$

$$\|Au - Au_n\| < \frac{\varepsilon}{K}.$$

Now from condition C, it follows that $\|u-u_n\|<\varepsilon$, if $n \geq n_0$ i.e. that $u_n \to u$.

We suppose that the conditions of lemma 2 are satisfied. Let u_n be the approximate solution of the equation $Au=f$. Inequality (6) gives a simple estimate of the error, i.e.

$$\|u_n-u\| \leq K\|A(u_n-u)\| = K\|Au_n-Au\|,$$

or

$$\|u_n-u\| \leq K\|Au_n-f\|. \tag{7}$$

If u_n is found by the method of least squares then $Au_n \to f$ as $n \to \infty$ and formula (7) thus allows us to estimate the error of the approximate solution.

NOTE: In practice a more convenient and rather different formulation of the method of least squares is used which is essentially equivalent to that given above.

Suppose that we require to find one or several functions of a certain number of independent variables and let these functions, which we denote by $u_1, u_2, \ldots, u_m$ satisfy two systems of linear equations: one homogeneous

$$K_j(u_1, u_2, \ldots, u_m) = 0, \quad j = 1, 2, \ldots, m \tag{*}$$

and the other non-homogeneous

$$G_j(u_1, u_2, \ldots, u_m)=b_j, \quad j=1, 2, \ldots, m.$$

We shall consider the sets $(u_1, u_2, \ldots, u_m)$ and $(b_1, b_2, \ldots, b_m)$ as elements u and b of some Hilbert space H. Then we can consider the sets of operators K_j and G_j as operators K and G in this same space.

How to choose the space H remains arbitrary to a considerable extent. We can use this arbitrariness to simplify the calculation or to obtain more precise results.

The method of least squares consists of the following:

(a) we construct a sequence $\{\varphi^{(k)}\}=\{(\varphi_1^{(k)}, \varphi_2^{(k)}, \ldots, \varphi_m^{(k)})\}$ $k=1, 2, \ldots$ of functions satisfying the homogeneous system (*);

(b) we construct an approximate solution of the problem in the form:

$$u^{(n)}=\sum_{k=1}^{n} a_k\varphi^{(k)}, \text{ or, specifically } u_j^{(n)}=\sum_{k=1}^{n} a_k\varphi_j^{(k)},$$

where the a_k are constants. We determine these constants from the condition that $\|Gu^{(n)}-b\|^2$ is a minimum,

or, what is the same thing, from the system of linear algebraic equations

$$\sum_{k=1}^{n} A_{jk}a_k=(b, G\varphi^{(j)}), \text{ where } A_{jk}=(G\varphi,^{(k)} G\varphi^{(j)}).$$

We now pause to establish the close connexion which exists between the method of least squares and the energy method.

We apply the operator A^*, the conjugate to A, to both sides of equation (1). (The operator A^*, conjugate to the given operator A is defined by the identity

$$(Au, v)=(u, A^*v).$$

For greater detail on the properties of the conjugate operator see e.g. V.I. Smirnov [3].) We shall prove that *the equation so obtained*

$$A^*Au=A^*f \tag{8}$$

is equivalent to equation (1) *if the latter is soluble* (cf. the author's paper [2]). We assume of course, that f belongs to the field of definition of the operator A^*. Let u' satisfy equation (1). Obviously it also satisfies equation (8), so that

$$A^*Au' = A^*f. \tag{9}$$

Let some further element u'' also satisfy equation (8) so that the identity

$$A^*Au'' = A^*f \tag{10}$$

s valid.

We shall prove that u'' satisfies equation (1). This will establish the validity of our assertion. We use the notation $u''-u'=v$. By subtracting (9) and (10) we find that $A^*Av=0$. We multiply this scalarly by v:

$$0=(A^*Av,\ v)=(Av,\ Av)=\|Av\|^2.$$

Hence $Av=A(u''-u')=0$ or $Au''=Au'=f$, which was to be proved.

It is readily further seen that *if condition C of lemma* 2 *is satisfied then the operator* A^*Au *is positive-bounded-below*. In fact

$$(A^*Au,\ u)=(Au,\ Au)=\|Au\|^2\geq\frac{1}{K^2}\|u\|^2.$$

We shall assume that the field of definition of the operator A^*A is dense in H, and that conditions B and C of lemma 2 are satisfied. Since equation (1) is soluble (condition B) then, as has been proved, it is equivalent to equation (8) and it is possible to find the solution of the latter. By virtue of condition C, the operator A^*A is positive-bounded-below. By theorem 1 § 9, finding the solution of equation (8) is equivalent to the problem of the minimum of the functional $\Phi(u)=(A^*Au,\ u)-(u,\ A^*f)-(A^*f,\ u)$. But, obviously,

$$\Phi(u)=(Au,\ Au)-(Au,\ f)-(f,\ Au)=(Au-f,\ Au-f)-(f,\ f),$$

or

$$\Phi(u)=\|Au-f\|^2-\|f\|^2.$$

The functionals $\Phi(u)$ and $\|Au-f\|^2$ differ by a constant term and the problems of finding minima for both functionals are the same. Thus, *under the conditions of lemma* 2, *the application of the method of least squares to equation* (1) *is equivalent to the application of the energy method to equation* (8).

Conversely, if the operator A is positive-bounded-below:

$$(Au,\ u)\geq\gamma^2\|u\|^2,\qquad \gamma\geq 0.$$

It is possible to apply the energy method to equation (1) by replacing the solving of this equation by finding the minimum of the functional

$$F(u)=(Au,\ u)-(u,\ f)-(f,\ u).$$

We construct the positive-bounded-below operator B whose square equals A, $B^2=A$.

In general the construction of the operator B is complicated and reqiures the application of the spectral theory of operators (see, e.g. V.I.

Smirnov [3]). It is comparatively simple to construct the operator B if the spectrum of operator A is discrete. Denoting the eigenvalues and eigenelements of operator A by λ_n and φ_n we have

$$Bu=\sum_{n=1}^{\infty}\gamma(\lambda_n)(u,\ \varphi_n)\varphi_n.$$

The operator B^{-1} exists and is bounded. Equation (1) is equivalent to the equation

$$Bu=B^{-1}f, \tag{11}$$

obtained from (1) by applying operator B^{-1} to both sides. Further

$$\begin{aligned}F(u)&=(B^2u,\ u)-(u,\ BB^{-1}f)-(BB^{-1}f,\ u)\\&=(Bu,\ Bu)-(Bu,\ B^{-1}f)-(B^{-1}f,\ Bu)\\&=(Bu-B^{-1}f,\ Bu-B^{-1}f)-(B^{-1}f,\ B^{-1}f),\end{aligned}$$

or

$$F(u)=\|Bu-B^{-1}f\|^2-\|B^{-1}f\|^2.$$

The functionals $F(u)$ and $\|Bu-B^{-1}f\|^2$ differ only by a constant term and the problems of finding minima for these functionals are the same. Hence it follows that, *in the case of a positive-bounded-below operator A, the application of the energy method to equation* (1) *is equivalent to the application of the method of least squares to equation* (11).

If the operator A is positive-bounded-below and u_n is an approximate solution of the equation $Au=f$ constructed by the Ritz method, then if $B^2=A$, $\|Bu_n-B^{-1}f\|\to 0$. This follows directly from the equivalence of the two methods just established and from the course of the proof of lemma 2. We recall once again that the relation $\|Au_n-f\|\to 0$ does not hold in general if u_n is constructed by the Ritz method.

We suppose once again that the operator Au is positive-bounded-below. We recall (cf. § 46) that the solution of equation (1) in this case is equivalent to the problem of finding the minimum of the functional

$$F(u)=(Au,\ u)-(u,\ f)-(f,\ u).$$

If we introduce a Hilbert space H_0 (cf. § 46) with the scalar product

$$[u,\ v]=(Au,\ v),$$

then, as we have seen,

$$F(u)=\|u-u_0\|^2-\|u_0\|^2,$$

where u_0 is the solution (1). Now let u_n be a solution obtained by the method of least squares. It is readily seen that *the*

sequence $\{u_n\}$ *tends to* u_0 *also in the metric of* H_0. (In other words, the sequence $\{u_n\}$ is minimizing for the functional $F(u)$.) In fact

$$\|u_n-u_0\|^2=(Au_n-Au_0,\ u_n-u_0)=(Au_n-f,\ u_n-u_0).$$

By Cauchy's inequality

$$\|u_n-u_0\|^2\leq\|Au_n-f\|\ \|u_n-u_0\|.$$

Further, by inequality (6),

$$\|u_n-u_0\|\leq K\|Au_n-Au_0\|=K\|Au_n-f\|,$$

and, consequently,

$$\|u_n-u_0\|^2\leq K\|Au_n-f\|^2, \tag{12}$$

whence it follows that

$$\|u_n-u_0\|^2\to 0.$$

Obviously formula (12) allows us to estimate the error of the approximate solution constructed by the method of least squares.

The quantity $\|Au_n-f\|^2$ can be evaluated in the following way. We have

$$\|Au_n-f\|^2=(Au_n-f,\ Au_n-f)=\left(\sum_{k=1}^{n}a_kA\varphi_k-f,\ \sum_{m=1}^{n}a_mA\varphi_m-f\right)$$

$$=\sum_{k,\,m=1}^{n}a_k\bar{a}_m(A\varphi_k,\ A\varphi_m)-\sum_{k=1}^{n}a_k(A\varphi_k,f)-\sum_{m=1}^{n}\bar{a}_m(f,\ A\varphi_m)+(f,\ f). \tag{13}$$

The coefficients a_k satisfy system (5)

$$\sum_{k=1}^{n}a_k(A\varphi_k,\ A\varphi_m)=(f,\ A\varphi_m).$$

Multiplying by $\bar{a}_m$ and summing over m, we obtain

$$\sum_{k,\,m=1}^{n}a_k\bar{a}_m(A\varphi_k,\ A\varphi_m)=\sum_{m=1}^{n}\bar{a}_m(f,\ A\varphi_m).$$

Substituting this in (13) we obtain

$$\|Au_n - f\|^2 = \|f\|^2 - \sum_{k=1}^{n} a_k (A\varphi_k, f)$$

$$= \|f\|^2 - \left(\sum_{k=1}^{n} a_k A\varphi_k, f \right) = \|f\|^2 - (Au_n, f). \tag{14}$$

The error of an approximate solution constructed by the method of least squares is estimated by the inequalities

$$\|u_n - u_0\|^2 \leq K^2 \{\|f\|^2 - (Au_n, f)\}, \tag{15}$$

$$\|u_n - u_0\|^2 \leq K \{\|f\|^2 - (Au_n, f)\}. \tag{16}$$

(We recall that u_0 denotes the solution of equation (1). Formula (16) only holds if the operator A is positive-bounded-below.)

It should be noted that, in the metric of the space H_1, the method of least squares converges more slowly than the Ritz method. In fact, we shall start from the same coordinate elements $\{\varphi_n\}$. Let

$$u_n = \sum_{k=1}^{n} a_k \varphi_k, \quad v_n = \sum_{k=1}^{n} b_k \varphi_k,$$

where the coefficients a_k are determined by the method of least squares and the coefficients b_k, by the Ritz method. Then, from the definition of the Ritz method,

$$F(u_n) \geq F(v_n); \quad \|u_n - u_0\| \geq \|v_n - u_0\|.$$

The last inequality demonstrates the slowness of convergence of the sequence $\{u_n\}$ as compared with the sequence $\{v_n\}$. The advantage of the method of least squares, however, consists of the fact that the sequence $\{u_n\}$ may converge uniformly (sometimes together with some of the derivatives) in cases where this is not possible for the sequence $\{v_n\}$.

§ 84. Application to integral equations

The method of least squares proves useful when applied to various classes of linear equations in Hilbert space. Let A be a linear bounded operator in the Hilbert space H such that

the bounded inverse operator A^{-1} exists. We consider the equation

$$Au=f, \tag{1}$$

where u is the required, and f the given element of the space. In using the method of least squares, we take a sequence of linearly independent elements $\{\varphi_n\}$ which is complete in H and we shall look for an approximate solution of equation (1) in the form

$$u_n=\sum_{k=1}^{n}a_k\varphi_k. \tag{*}$$

We now verify that conditions $A-C$ of lemma 2 § 83 are satisfied. Condition A follows from the fact that the sequence $\{\varphi_n\}$ is complete in H and the operator A is bounded; conditions B and C follow from the boundedness of the operator A^{-1}. On the basis of lemma 2 § 83, $u_n \to u$ where u is the solution of equation (1).

Now suppose that the operator A has no inverse but there exists a bounded regularizing operator M, i.e. one such that $MA=E+T$, where E is the identity operator and T is a completely continuous operator. In this case for the solubility of equation (1), it is necessary and sufficient that f be orthogonal to all the solutions of the homogeneous conjugate equation $A^*u=0$ (cf. the author's paper [5]). Further, if we denote by H_1^* the linear set of elements from H which are orthogonal to all the solutions of the equation $A^*u=0$, and by H_1, the linear set of elements which are orthogonal to all the solutions of the equation $Au=0$, and we consider the operator A only in H_1, then the inverse operator A^{-1} exists and is bounded in H_1^*. (We can readily convince ourselves that H_1 and H_1^* are complete Hilbert spaces [in the sense of § 43].) Our method for the approximate solution can also be used in the case where the homogeneous equation $Au=0$ has only a finite number of linearly independent solutions $u^{(1)}$, $u^{(2)}, \ldots, u^{(r)}$ and these solutions are known. Assuming that $f \in H_1^*$, we shall seek the approximate solution in H; for this

it suffices to represent it in the form (*), but to make the coefficients satisfy the equations:

$$\sum_{k=1}^{n} a_k(\varphi_k, u^{(j)}) = 0, \quad j = 1, 2, \ldots, r.$$

With the aid of these equations, it is possible to eliminate some of the coefficients a_k and we determine the remainder from the condition $\|Au_n - f\|^2 = \min$, which leads to system (5) § 83.

We now apply the method of approximate solution given here to Fredholm integral equations (cf. M. Kravchuk [2]) whose kernels satisfy the condition:

$$\int_{\Omega}\int_{\Omega} |K^2(P,Q)| \, d\Omega_P d\Omega_Q,$$

and also to singular integral equations (involving the principal value of the integral) provided that the integral is over a smooth closed curve. In fact such a singular operator is bounded in the Hilbert space of functions with summable squares and has a bounded regularizing operator. (The proof of these propositions can be found in the author's paper [5].)

In its application to integral equations it is possible in certain cases to alter the method so that the convergence becomes uniform.

Suppose we are given the Fredholm equation

$$Au = u(P) - \int_{\Omega} K(P,Q)u(Q)d\Omega_Q = f(P), \tag{2}$$

where $P(x_1, x_2, \ldots, x_m)$ and $Q(s_1, s_2, \ldots, s_m)$ are points in the m-dimensional domain Ω. Let $f(P)$ be $\left[\frac{m}{2}\right]+1$ times differentiable with respect to all the coordinates $x_1, x_2, \ldots, x_m$ and let the structure of the equation be such that the solution $u(P)$ is also $\left[\frac{m}{2}\right]+1$ times differentiable, and, finally, let the operator on the left-hand side of (2) transform a function

which is differentiable a sufficient number of times into a function differentiable $\left[\frac{m}{2}\right]+1$ times. As the $\varphi_k(P)$ we take a sequence of functions which are differentiable a sufficient number of times, e.g. polynomials, and as before we put $u_n(P)=\sum_{k=1}^{n} a_k\varphi_k(P)$. The coefficients a_k however will be determined from the condition

$$\|Au_n-f\|^2+\sum_{k=1}^{[\frac{m}{2}]+1}\sum_{j_1+\ldots j_m=k}\left\|\frac{\partial^k(Au_n-f)}{\partial x_1^{j_1}\ldots\partial x_n^{j_n}}\right\|^2=\min. \tag{3}$$

(We define the norm in the usual way: $\|u\|^2=\int_{\Omega}u(P)^2\mathrm{d}\Omega$.)

Assuming that a solution of equation (2) exists, we readily find that

$$\|Au_n-f\|\xrightarrow[n\to\infty]{}0,\quad \left\|\frac{\partial^k(Au_n-f)}{\partial x_1^{j_1}\ldots\partial x_n^{j_n}}\right\|\longrightarrow 0,$$

$$k=1,\ 2,\ \ldots,\left[\frac{m}{2}\right]+1.$$

According to the well-known Parseval formula $Au_n\to f$ uniformly in Ω. We further suppose that

$$\int_{\Omega}|K^2(P,Q)|\,\mathrm{d}\Omega<C^2,\ C=\mathrm{const}.$$

We use the notation $Au_n=f_n$. Then

$$u(P)-u_n(P)=\int_{\Omega}K(P,Q)[u(Q)-u_n(Q)]\mathrm{d}\Omega_Q+f(P)-f_n(P),$$

whence, by virtue of Cauchy's inequality

$$|u(P)-u_n(P)|<C\|u-u_n\|+|f(P)-f_n(P)|.$$

As proved above (cf. § 83), it follows from $\|Au_n-f\|\to 0$ that $\|u-u_n\|\to 0$ and from the last inequality it is clear that $u_n(P)\to u(P)$ uniformly within Ω. By introducing derivatives

of higher order into the left-hand side of (3), we can also convince ourselves of the uniform convergence of the derivatives under known conditions.

§ 85. Application to boundary value problems with homogeneous boundary conditions

The method of least squares has frequently been applied to problems of this type in the works of N. M. Krylov and his successors. Questions connected with this are discussed in detail in a monograph by M. Kravchuk [1] to which we refer the reader; also given in the monograph is an extensive list of the works of N. M. Krylov's school on the method of least squares and on the so called "method of moments". In the present section we shall show, using the Poisson equation as an example, the special features of the application of the method of least squares to problems of the type mentioned.

Thus, let us integrate the Poisson equation

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y) \tag{1}$$

in the domain Ω under the boundary conditions

$$u|_S = 0. \tag{2}$$

We assume that the boundary S is sufficiently smooth.

We select a sequence of coordinate functions $\{\varphi_n(x, y)\}$ which are differentiable a sufficient number of times in $\overline{\Omega} = \Omega + S$ and satisfy condition (2). As usual we look for an approximate solution in the form

$$u_n = \sum_{k=1}^{n} a_k \varphi_k$$

and we determine the coefficients a_k from the condition that

$$\int_{\Omega} [\nabla^2 u_n - f]^2 \mathrm{d}\Omega \text{ is to be a minimum.}$$

(We assume all the given functions to be real.)

As in the general case it is readily proved that

$$\|\nabla^2 u_n - f\|^2 = \int_\Omega [\nabla^2 u_n - f]^2 \mathrm{d}\Omega \xrightarrow[n\to\infty]{} 0.$$

We denote Green's function for our problem by $G(x, y; \xi, \eta)$. Then, since $u|_s = u_n|_s = 0$,

$$u(x, y) - u_n(x, y) = \int_\Omega G(x, y; \xi, \eta)(\nabla^2 u - \nabla^2 u_n)\mathrm{d}\xi\mathrm{d}\eta$$

$$= \int_\Omega G(x, y; \xi, \eta)(f - \nabla^2 u_n)\mathrm{d}\xi\mathrm{d}\eta. \tag{3}$$

We apply Cauchy's inequality to the last integral

$$[u(x, y) - u_n(x, y)]^2 \leq \int_\Omega G^2(x, y; \xi, \eta)\mathrm{d}\xi\mathrm{d}\eta \int_\Omega (f - \nabla^2 u_n)^2 \mathrm{d}\Omega. \tag{4}$$

The function $G(x, y; \xi, \eta)$ is only logarithmically infinite and so the first integral is bounded; the second integral tends to zero and inequality (4) shows that $u_n(x, y) \to u(x, y)$ *uniformly* in $\overline{\Omega}$. From the reasoning given at the end of the previous section it follows that $\dfrac{\partial u_n}{\partial x} \to \dfrac{\partial u}{\partial x}$, $\dfrac{\partial u_n}{\partial y} \to \dfrac{\partial u}{\partial y}$ in the mean (cf. the author's paper [6]) in the domain Ω.

These results show that the method of least squares, in the case under discussion, is more convergent than the Ritz method, although the convergence in the mean of the derivatives is somewhat slower.

§ 86. Auxiliary theorems of the theory of analytic functions

For the rest of the chapter, we shall restrict ourselves for simplicity to the case of a finite singly-connected domain in the plane of the complex variable $z = x + iy$. The extension of the results to the case of a multiply-connected or infinite domain encounters no essential difficulties but the treatment becomes far more involved (cf. the author's paper [6]).

As in all the foregoing, we shall denote the domain considered by Ω and its boundary by S: the closed domain $\Omega+S$ will be denoted as before by $\overline{\Omega}$. We shall always locate the coordinate origin within Ω. We shall always consider the curve S to be sufficiently smooth.

The following theorem due to Walsh [1] will be of great importance in what follows.

THEOREM 1. *Let the function $f(z)$ be holomorphic in Ω and continuous in $\overline{\Omega}$. For any $\varepsilon>0$ it is possible to find a polynomial $\varphi(z)$ such that the inequality*

$$|f(z)-\varphi(z)|<\varepsilon \tag{1}$$

is satisfied within $\overline{\Omega}$.

We note two corollaries of theorem 1.

COROLLARY 1. If $f(z)$ is holomorphic within Ω and continuous in Ω together with its derivatives up to order r inclusive, then for any number $\varepsilon>0$ it is possible to find a polynomial $\varphi(z)$ which satisfies the inequalities

$$|f(z)-\varphi(z)|<\varepsilon, \quad |f'(z)-\varphi'(z)|<\varepsilon, \quad \ldots$$
$$|f^{(r)}(z)-\varphi^{(r)}(z)|<\varepsilon \tag{2}$$

within $\overline{\Omega}$.

By Walsh's theorem it is possible to construct a polynomial $\varphi(z)$ which satisfies the inequality

$$|f^{(r)}(z)-\varphi(z)|<\frac{\varepsilon}{A}.$$

We shall determine the constant A below. We shall find the polynomial $f^*(z)$ from the conditions $f^{*(r)}(z)=\varphi(z)$, $f^{*(j)}(z_0)=f^{(j)}(z_0)$, $j=0,1,2,\ldots,r-1$ (where z_0 is some fixed point within Ω). By integrating we find:

$$|f^{(r-1)}(z)-f^{*(r-1)}(z)|=\left|\int_{z_0}[f^{(r)}(z)-\varphi(z)]\mathrm{d}r\right|<\frac{\varepsilon d}{A},$$

where d is the maximum of the shortest paths between z_0 and points on S lying within $\overline{\Omega}$.

Similarly

$$|f^{(r-2)}(z)-f^{*(r-2)}(z)|<\frac{\varepsilon d^2}{A},\ \ldots,\ |f(z)-f^*(z)|<\frac{\varepsilon d^r}{A}.$$

It suffices to take $A=1$ if $d\leq 1$, and $A=d^r$ if $d>1$ and inequalities (2) are proved.

COROLLARY 2. Under the conditions of corollary 1, it is possible for a given number $\varepsilon>0$, to construct a polynomial $\varphi(z)$ such that

$$\sum_{j=0}^{r}\int_S |f^{(j)}(z)-\varphi^{(j)}(z)|^2 ds<\varepsilon^2,\quad ds=|dz|.$$

The proof is obvious.

Below we shall need to consider the space $L_2^{(r)}(S)$ whose elements are functions defined on the curve S. Considered as functions of the arc along S, they are continuous together with their derivatives with respect to the arc length up to the order $r-1$ inclusive, and their generalized derivatives of order r are quadratically summable along S. The scalar product in $L_2^{(r)}(S)$ is defined by the formula

$$(\varphi,\psi)_r=\sum_{j=0}^{r}\int_S \varphi^{(j)}(z)\overline{\psi^{(j)}(z)}\,ds;$$

hence follows the formula for the norm in $L_2^{(r)}(S)$:

$$\|\varphi\|_r^2=\sum_{j=0}^{r}\int_S |\varphi^{(j)}(z)|^2 ds.$$

The following lemma will be important in what follows.

LEMMA 1. *Let $\varphi\in L_2(S)$ and let z be a point on the curve S. The integral*

$$\psi(z)=\frac{1}{2\pi i}\int_S \frac{\varphi(\zeta)}{\zeta-z}\,d\zeta,$$

taken in the sense of its principal Cauchy value, is a bounded operator in $L_2(S)$.

From lemma 1 there follows the existence of a constant K depending only upon the curve S for which the inequality $\|\psi\|\leq K\|\varphi\|$ is valid.

The proof of lemma 1 can be found in the author's paper [5].

In the sequel we shall say that some sequence converges uniformly within Ω if this sequence converges uniformly in any closed domain which lies entirely with Ω.

THEOREM 2. *If the functions $f_n(z)$ are holomorphic within Ω and are represented by Cauchy's formula and if $\|f_n\| \xrightarrow[n\to\infty]{} 0$, then these functions tend to zero uniformly within Ω.*

It suffices, e.g. that the functions $f_n(z)$ which are holomorphic in Ω be continuous in the closed domain Ω.

We have

$$f_n(z) = \frac{1}{2\pi i} \int_{\Omega} \frac{f_n(\zeta)}{\zeta - z} \, d\zeta.$$

Let z lie within the closed domain $\overline{\Omega}' \subset \Omega$. We denote by d, $d > 0$, the greatest distance between the boundaries of the domains $\overline{\Omega}'$ and Ω and by l the length of the curve S. Then

$$|f_n(z)| \geq \frac{1}{2\pi d} \int_{S} |f_n(\zeta)| \, |d\zeta|,$$

or, applying Cauchy's inequality to the last integral,

$$|f_n(z)| \leq \frac{\sqrt{l}}{2\pi d} \|f_n\| \xrightarrow[n\to\infty]{} 0.$$

COROLLARY. Under the conditions of theorem 2, $f_n(z) \to 0$ at any internal point of Ω.

THEOREM 3. *If $f_n(z)$ is holomorphic within Ω, $f_n^{(r)}(z)$ is represented by Cauchy's formula and $\|f_n\|_r \to 0$, then the functions $f_n(z)$ tend to zero uniformly within $\overline{\Omega}$ together with their derivatives of order up to $r-1$ inclusive.*

For simplicity we restrict ourselves to the case $r = 1$. The general case can be discussed analogously. Firstly we note that $\|f_n'(z)\| \to 0$ if $\|f_n\|_1 \to 0$. Further, by integrating Cauchy's formula for the derivative, we obtain

$$f_n(z) - f_n(z_0) = \frac{1}{2\pi i} \int_{S} f_n'(\zeta) \ln \frac{\zeta - z_0}{\zeta - z} \, d\zeta,$$

where z_0 is a point within Ω. Now by Cauchy's inequality

$$|f_n(z)-f_n(z_0)|^2 \leq \frac{\|f_n'\|^2}{4\pi^2}\int_S \left|\ln\frac{\zeta-z_0}{\zeta-z}\right|^2 |\mathrm{d}\zeta|.$$

If z lies within the closed domain $\bar{\Omega}$, the last integral is bounded and $f_n(z)-f_n(z_0)\to 0$ uniformly in $\bar{\Omega}$. But $f_n(z_0)\to 0$ (corollary of theorem 2) and our theorem is proved.

THEOREM 4. *Let the function $f(z)=u(x, y)+iv(x, y)$ be holomorphic in Ω and continuous in $\bar{\Omega}$ and let $v(0, 0)=0$. Then there exists a constant K, which depends only upon the domain Ω such that*

$$\|f\| \leq K\|u\|. \tag{3}$$

We transform Ω conformally into the unit circle in the t-plane; let $t=\chi(z)$ be the function which realizes this transformation. We introduce the notation

$$t=re^{i\vartheta}; \quad u(x, y)=U(r, \vartheta); \quad f(z)=F(t).$$

We further denote the circle $|t|=1$ by γ. Within Ω the function $F(t)$ can be expanded in a Taylor series.

$$F(t)=\sum_{n=1}^{\infty} a_n t^n.$$

We further use the notation $a_n=\alpha_n+i\beta_n$. We perform the conformal transformation so that the point $z=0$ transforms to the point $t=0$. Then Im $(a_0)=0$. A simple calculation gives

$$\int_\gamma U^2(1, \vartheta)\mathrm{d}\vartheta = 2\pi a_0^2+\pi\sum_{n=1}^{\infty}|a_n|^2; \quad \int_\gamma |F^2(e^{i\vartheta})|\,\mathrm{d}\vartheta = 2\pi\sum_{n=0}^{\infty}|a_n|^2.$$

Hence

$$\int_\gamma |F^2(e^{i\vartheta})|\,\mathrm{d}\vartheta \leq 2\int_\gamma U^2(1, \vartheta)\mathrm{d}\vartheta. \tag{4}$$

Reverting to the variable z, we obtain

$$\int_S |f^2(z)|\,|\chi'(z)|\,\mathrm{d}s \leq 2\int_S u^2|\chi'(z)|\,\mathrm{d}s.$$

Since the curve S is smooth, $\chi'(z)$ is bounded in modulus above and below. Let $A \leq |\chi'(z)| \leq B$. From (4) it follows that

$$\|f\| \leq K\|u\|, \quad K = \sqrt{\left(\frac{2B}{A}\right)}.$$

With suitable supplementary conditions imposed upon $f(z)$ we can readily obtain the estimate

$$\|f\|_r \leq K_r\|u\|_r. \tag{5}$$

In conclusion, we state that unique approximation of analytic functions by polynomials is impossible. An example of approximation by functions of another type will be given in § 88.

§ 87. The Dirichlet and Neumann problems

We consider the Dirichlet problem for a singly-connected finite domain Ω.

We shall denote the required harmonic function by $u(x, y)$, its values on the boundary S of the domain by $u(\zeta)$ and the holomorphic function in Ω whose real part coincides with $u(x, y)$ by $f(z)$. The function $f(z)$ is defined to within a purely imaginary term which we define by the condition.

$$\operatorname{Im}\{f(0)\} = 0. \tag{1}$$

In following the method of least squares, we shall look for an approximate solution of the Dirichlet problem in the form

$$f_n(z) = u_n + iv_n = \sum_{k=0}^{n} a_k z^k. \tag{2}$$

In order to satisfy condition (1), we stipulate that

$$\operatorname{Im}(a_0) = 0. \tag{3}$$

Now we determine the coefficients in the expression for $f_n(z)$ from the condition

$$\|u - u_n\|^2 = \min. \tag{4}$$

We now investigate the convergence of the approximate solutions. Condition A of lemma 2 (§ 83) is a simple corollary of Walsh's theorem. Further, condition B follows from the existence of a solution of the Dirichlet problem.
Condition C is also obviously satisfied, since in our case $Au \equiv u$. By virtue of lemma 2 § 83 $\|u-u_n\| \to 0$.

We now prove that $f_n(z) \to f(z)$ *uniformly within* Ω. In fact from estimate (3) § 86 it follows that $\|f_n(z)-f(z)\| \to 0$, and from theorem 2 § 86, that $f_n(z)-f(z) \to 0$ uniformly within Ω. (This result with certain additional restrictions was given by M. Piccone [1] who only considered the case of a singly-connected domain.)
NOTES: 1. The above reasoning obviously holds if, instead of (2), we take a function of the form:

$$f_n(z) = \sum_{k=1}^{n} a_k \omega_k(z),$$

provided that $\{\omega_k(z)\}$ is a complete system of functions holomorphic in Ω and continuous in $\bar{\Omega}$.
2. It is possible to alter the process so that uniform convergence occurs in $\bar{\Omega}$. For this it suffices that the function $f_n(z)$ be subject to the condition

$$\|u-u_n\|_1 = \min. \qquad (4_1)$$

In this case the evaluation of the coefficients becomes rather more complex.
3. If we stipulate that the quantity $\|u-u_n\|_r$ shall be a minimum, then in $\bar{\Omega}$ there will occur the uniform convergence of the derivatives $f_n^{(j)}(z)$, $j=1, 2, \ldots, r-1$ to the corresponding derivatives $f^{(j)}(z)$.

With the purpose of illustrating the method, we shall consider the Dirichlet problem for a circle. We take the radius of the circle equal to unity and place the coordinate origin at the centre of the circle. We expand the function $u(\zeta)$ specified on S into the Fourier series:

$$u(\zeta) = A_0 + \sum_{k=1}^{\infty} (A_k \cos k\theta + B_k \sin k\theta), \quad \zeta = e^{i\theta}.$$

We shall look for the function $f_n(z)$ in the form of a polynomial; this is equivalent to putting $\omega_k(z)=z^k=r^k e^{ik\theta}$

Putting

$$f_n(z)=a_0+\sum_{k=1}^{n}(a_k-ib_k)z^k, \quad \mathrm{Im}(a_0)=0,$$

we find

$$a_0=A_0, \quad a_k=A_k, \quad b_k=B_k.$$

This gives for the function $f_n(z)$, the power series of the function $f(z)$ terminating at term n.

The Neumann problem consists of seeking a function $f(z)=u(x, y)+iv(x, y)$ which is holomorphic in Ω and satisfies the boundary condition

$$\left.\frac{\partial u}{\partial n}\right|_s=\psi(s). \tag{5}$$

The function $f(z)$ is defined to within an arbitrary constant term. We define this by the condition

$$f(0)=0. \tag{6}$$

THEOREM 1. *If condition* (6) *is satisfied then there exists a constant* K, *depending only upon the domain* Ω, *such that in the closed domain* $\bar{\Omega}$ *the inequality*

$$|f(z)|<K\left\|\frac{\partial u}{\partial n}\right\| \tag{7}$$

is satisfied.

We suppose that $f(z)$ and $f'(z)$ are continuous in $\bar{\Omega}$.

We transform Ω conformally into the unit circle with the aid of the function $t=\chi(z)$, $\chi(0)=0$. We use the same notation as in § 86,

$$t=re^{i\vartheta}, \quad f(z)=F(t), \quad u(x, y)=U(r, \vartheta).$$

We have $F(0)=f(0)=0$, therefore

$$F(t)=\sum_{n=1}^{\infty}a_n t^n, \quad a_n=\alpha_n+i\beta_n,$$

$$U(r, \vartheta)=\sum_{n=1}^{\infty}r^n(\alpha_n\cos n\vartheta-\beta_n\sin n\vartheta)$$

and

$$\int_0^{2\pi} \left(\frac{\partial U}{\partial r}\right)^2_{r=1} \mathrm{d}\vartheta = \pi \sum_{n=1}^{\infty} n^2 |a_n|^2,$$

$$|F(t)|^2 \leq \left(\sum_{n=1}^{\infty} |a_n|\right)^2$$

$$= \left(\sum_{n=1}^{\infty} n|a_n| \frac{1}{n}\right)^2 \leq \sum_{n=1}^{\infty} n^2 |a_n|^2 \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} \sum_{n=1}^{\infty} n^2 |a_n|^2.$$

Hence

$$|F(t)|^2 \leq \frac{\pi}{6} \int_0^{2\pi} \left(\frac{\partial U}{\partial r}\right)^2_{r=1} \mathrm{d}\vartheta.$$

Reverting to the variable z, and noting that because the transformation is conformal,

$$\frac{\partial U}{\partial r} = \frac{\partial u}{\partial n} \frac{1}{|\chi'(z)|}, \quad \mathrm{d}\vartheta = |\chi'(z)| \, \mathrm{d}s,$$

we obtain

$$|f(z)|^2 \leq \frac{\pi}{6} \int_S \left(\frac{\partial u}{\partial n}\right)^2 \frac{\mathrm{d}s}{|\chi'(z)|}.$$

Finally $|\chi'(z)| \geq A > 0$, $A = \text{const.}$, since the curve S is smooth. Hence

$$|f(z)| \leq K \left\| \frac{\partial u}{\partial n} \right\|, \quad K = \sqrt{\frac{\pi}{6A}}.$$

According to the method of least squares, an approximate solution of the Neumann problem can be found in the form

$$f_n(z) = \sum_{k=1}^{n} a_k z^k.$$

As was stated previously, it is possible to take instead of rational fractions linear combinations of the terms of any

sequence, complete in Ω, of functions which are holomorphic in Ω and continuous in $\bar{\Omega}$.

The coefficients of the expression for $f_n(z)$ are determined from the condition that

$$\left\| \frac{\partial u}{\partial n} - \frac{\partial u_n}{\partial n} \right\|^2 \text{ is a minimum.} \tag{8}$$

This gives a system of linear equations which has a unique solution. In fact, this follows from the fact that the condition of lemma 1 § 83 is satisfied in the Neumann problem, which in turn follows from theorem 1.

We now prove that $f_n(z) \to f(z)$ *uniformly* within $\bar{\Omega}$. By corollary 2 § 86, it is possible to find a polynomial $\varphi_{n_0}(z)$ such that $\|f'(z) - \varphi_{n_0}(z)\| < \varepsilon$. Further

$$\left\| \frac{\partial u}{\partial n} - \frac{\partial u_{n_0}}{\partial n} \right\| = \|\mathrm{Re}\{[f'(z) - \varphi'_{n_0}(z)]e^{-i(n,x)}\}\| \leq \|f'(z) - \varphi'_{n_0}(z)\| < \varepsilon.$$

A fortiori, by virtue of (8) we have

$$\left\| \frac{\partial u}{\partial n} - \frac{\partial u_{n_0}}{\partial n} \right\| < \varepsilon,$$

if u_{n_0} is found by the method of least squares. Obviously when $n > n_0$

$$\left\| \frac{\partial u}{\partial n} - \frac{\partial u_n}{\partial n} \right\| \leq \left\| \frac{\partial u}{\partial n} - \frac{\partial u_{n_0}}{\partial n} \right\| < \varepsilon.$$

Hence it follows that

$$\left\| \frac{\partial u}{\partial n} - \frac{\partial u_n}{\partial n} \right\| \to 0.$$

and there follows from (7) the uniform convergence of $f_n(z)$ to $f(z)$ in the closed domain $\bar{\Omega}$.

All the conclusions of the present section still hold if we introduce the space $L_2(p; S)$, instead of $L_2(S)$ under the condition that the weight function $p(\sigma)$ is bounded above and below by positive numbers

$$0 < \alpha \leq p(\sigma) \leq \beta < \infty.$$

§ 88. The Dirichlet problem for an ellipse

Suppose we are required to determine a function $u(x, y)$, which is harmonic within the ellipse $x=a\cos t$, $y=b\sin t$, $0\leq t\leq 2\pi$, under the boundary condition

$$u|_S = u(\sigma) \tag{1}$$

where S is the edge of the ellipse.

As above, we shall consider $u(x, y)$ to be the real part of some analytic function $f(z)$:

$$u(x, y) = \mathrm{Re}\{f(z)\}, \quad z = x + iy. \tag{2}$$

We adopt the functions

$$\varphi_0(z) = 1, \ \varphi_k(z) = [z + \sqrt{(z^2 - c^2)}]^k + [z - \sqrt{(z^2 - c^2)}]^k, \ k > 1 \tag{3}$$

as the coordinate functions ($c = \sqrt{(a^2 - b^2)}$. As is known, the system forms a complete linear set of functions which are holomorphic within the ellipse and continuous along its edge. We shall look for an approximate solution in the form:

$$f_n(z) = \sum_{k=0}^{n} a_k \varphi_k(z), \quad \mathrm{Im}(a_0) = 0. \tag{4}$$

Denoting the length of arc of the ellipse by $d\sigma$, we have

$$d\sigma = \sqrt{(a^2 \sin^2 t + b^2 \cos^2 t)}\, dt.$$

As the fundamental Hilbert space for our problem, we take the space $L_2(p; S)$ where

$$p(\sigma) = \frac{1}{\sqrt{(a^2 \sin^2 t + b^2 \cos^2 t)}}.$$

Obviously in $L_2(p; S)$ the scalar product of the functions φ and ψ equals

$$(\varphi, \psi) = \int_0^{2\pi} \varphi(\sigma)\overline{\psi(\sigma)}\, dt. \tag{5}$$

Following the method of least squares, we determine the coefficients a_k from the condition

$$\left\| \mathrm{Re} \sum_{k=0}^{n} a_k \varphi_k(z) - u(\sigma) \right\|^2 = \min. \tag{6}$$

On the edge of the ellipse $z=a\cos t+ib\sin t$ and, consequently, $z^2-c^2=(b\cos t+ia\sin t)^2$. Hence

$$\varphi_k(z)=(a+b)^k e^{ikt}+(a-b)^k e^{-ikt}, \quad k>0.$$

We use the notation $a_k=\alpha_k-i\beta_k$. Then on the ellipse

$$\begin{aligned}\mathrm{Re}\{a_k\varphi_k(z)\}&=[(a+b)^k+(a-b)^k]\alpha_k\cos kt\\&+[(a+b)^k-(a-b)^k]\beta_k\sin kt, \quad k>0.\end{aligned} \tag{7}$$

Now it is clear from (6) that the coefficients in (7) are the Fourier coefficients of the function $u(\sigma)$; hence

$$\left.\begin{aligned}\alpha_0=a_0&=\frac{1}{2\pi}\int_0^{2\pi}u(\sigma)\,dt,\\ \alpha_k&=\frac{1}{\pi[(a+b)^k+(a-b)^k]}\int_0^{2\pi}u(\sigma)\cos kt\,dt,\\ \beta_k&=\frac{1}{\pi[(a+b)^k-(a-b)^k]}\int_0^{2\pi}u(\sigma)\sin kt\,dt.\end{aligned}\right\} \tag{8}$$

The coefficients α_0, α_k, β_k, as is clear from (8), do not depend upon the number n, therefore the limiting process as $n\to\infty$ is readily realized in (4) giving the exact solution of the problem

$$f(z)=a_0+\sum_{k=1}^{\infty}a_k[(z+\sqrt{(z^2-c^2)})^k+(z-\sqrt{(z^2-c^2)})^k];$$

the coefficients a_k being defined by formulae (8).

§ 89. The case of a piecewise smooth contour. The Dirichlet problem

Let the contour S which bounds the domain Ω be piecewise smooth; we denote by $\zeta_1, \zeta_2, \ldots, \zeta_r$ its angular points and by $\pi\alpha_1, \pi\alpha_2, \ldots, \pi\alpha_r$ the internal angles at the corresponding

points. We introduce the Hilbert space $L_2(p; S)$ where the weight function is

$$p(\sigma)=\prod_{k=1}^{r}|\zeta-\zeta_k|^{\frac{1}{\alpha_k}-1}. \tag{1}$$

The scalar product and norm in $L_2(p; S)$ are defined by the formulae

$$(\varphi, \psi)=\int_S p(\sigma)\varphi(\zeta)\overline{\psi(\zeta)}\,d\sigma;\ \|\varphi\|_S^2=\int_S p(\sigma)|\varphi^2(\zeta)|\,d\sigma. \tag{2}$$

LEMMA: *There is a set of polynomials, dense in the sense of metric* (2), *in the class of functions which are holomorphic in* Ω *and representable by Cauchy's formula, and the limiting values of the real parts of which belong to the space* $L_2(p; S)$.

In the proof we shall assume, as above, that Ω is singly-connected. We map Ω conformally onto the unit circle; let $t=\chi(z)$ be the transforming function. As is known,

$$\chi'(z)=\varkappa(z)\prod_{k=1}^{r}(z-\zeta_k)^{\frac{1}{\alpha_k}-1}, \tag{3}$$

where $\varkappa(z)$ is bounded both above and below in the closed domain $\bar{\Omega}=\Omega+S$ by positive numbers

$$0<A\leq\varkappa(z)\leq B<\infty. \tag{4}$$

We denote the circle $|t|=1$ by γ. We introduce the Hilbert space $L_2(\gamma)$ in which

$$(\varphi, \psi)_\gamma=\int_\gamma\varphi(t)\overline{\psi(t)}\,d\vartheta,\quad \|\varphi\|_\gamma^2=\int_\gamma|\varphi^2(t)|\,d\vartheta,\quad t=e^{i\vartheta}. \tag{5}$$

Let $u\in L_2(p; S)$. Then

$$\|u\|_\gamma^2=\int_\gamma|u|^2d\vartheta=\int_S|u^2|p(\sigma)|x(z)|\,d\sigma\leq B\|u\|_S^2<\infty. \tag{6}$$

Hence it follows that $u\in L_2(\gamma)$; putting $t=e^{i\vartheta}$, we find that u expands into a Fourier series which converges in the mean:

$$u=a_0+\sum_{k=1}^{\infty}{}'(\alpha_k\cos k\vartheta-\beta_k\sin k\vartheta).$$

We use the notation $a_k=\alpha_k+i\beta_k$ and, assuming the function u real, we put

$$F(t)=\alpha_0+i\beta_0+\sum_{k=1}^{\infty}a_kt^k,\quad F_n(t)=\alpha_0+i\beta_0+\sum_{k=1}^{n}a_kt^k,\ \beta_0=\text{const.}$$

The function $F(t)$ is representable by Cauchy's formula and its real part coincides with u on the circle γ. It is readily seen that

$$\|F-F_n\|_\gamma^2=2\pi\sum_{k=n+1}^{\infty}|a_k|^2 \tag{7}$$

and, consequently, for sufficiently large n, $\|F-F_n\|<\varepsilon$ where ε is an arbitrary positive number. We now put $F(t)=f(z)$, $F_n(t)=f_n(z)$. Formula (7) then transforms to the following:

$$\int_S|f(z)-f_n(z)|^2p(\sigma)|\varkappa(z)|\,d\sigma=2\pi\sum_{k=n+1}^{\infty}|a_k|^2<\varepsilon^2,$$

whence it readily follows that

$$\|f-f_n\|<\frac{\varepsilon}{\sqrt{(A)}}.$$

By Walsh's theorem the function $f_n(z)$ which is continuous in $\bar{\Omega}=\Omega+S$ and holomorphic in Ω can be approximated uniformly and, *a fortiori*, in the sense of the metric of $L_2(p;S)$ by the polynomial $f^*(z)$. But

$$\|f-f^*\|\le\|f-f_n\|_S+\|f_n-f^*\|_S<\frac{\varepsilon}{\sqrt{(A)}}+\|f_n-f^*\|_S$$

and since both terms on the right are arbitrarily small, $\|f-f^*\|$ is also arbitrarily small, which proves the lemma.

We now pass on to the Dirichlet problem. It is readily seen that estimate (3) § 86 remains valid in $L_2(p;S)$. In fact, it is valid in $L_2(\gamma)$. Thus if $f(z)=u+iv$ is holomorphic in Ω and $\operatorname{Im}\{f(0)\}=0$, then by inequality (3) § 86:

$$\|f\|_\gamma<K\|u\|_\gamma,\qquad\int_\gamma|f(z)|^2d\vartheta\le K\int_\gamma u^2d\vartheta$$

also $d\vartheta = p(\sigma)|\chi(z)|\,d\sigma$; substituting this in the last inequality and using (4), we obtain

$$\|f\|_S \geq K'\|u\|_S; \quad K' = K\sqrt{\left(\frac{B}{A}\right)}, \tag{8}$$

which is identical with estimate (3) § 83. To solve the Dirichlet problem we require that the expression

$$\|u - \mathrm{Re}\{f^*(z)\}\|_S^2 = \min,$$

where $f^*(z)$ is a polynomial. The lemma of the present section allows us to assert that condition A of § 83 is satisfied; condition B follows from the existence theorem for the Dirichlet problem, and condition C from inequality (8). Thus the convergence of the method of least squares in the metric of $L_2(p; S)$ is established.

§ 90. The mixed problem of potential theory

Here we shall consider the problem of finding the harmonic function $u(x, y)$ which is continuous in the closed domain $\bar{\Omega} = \Omega + S$, if, on some parts of the contour S, the values of the function $u(x, y)$ itself are specified and if on the others, the values of its normal derivative are specified.

We shall restrict ourselves to the case where Ω is finite and singly-connected. As always we shall consider the contour S to be sufficiently smooth. Let the points $\alpha_1, \alpha_2, \ldots, \alpha_{2m}$, $\alpha_{2m+1} = \alpha_1$ divide S into the arcs $\gamma_1, \gamma_2, \ldots, \gamma_{2m}$ so that arc γ_k starts and finishes at the points α_k and α_{k+1} respectively. We shall use the notation

$$\sum \gamma_{2k-1} = \lambda_1; \quad \sum \gamma_{2k} = \lambda_2.$$

Further let the values $u(x, y) = \varphi(\sigma)$ be specified on λ_1, and let the values $\dfrac{\partial u}{\partial n} = \psi(\sigma)$ be specified on λ_2. We further use the notation

$$R(\sigma) = \sqrt{\left(\prod_{k=1}^{m} |\zeta - \alpha_k|\right)}.$$

THEOREM 1. *If* $\varphi(\sigma)\equiv 0$, *then there exists a constant* K *which is independent of the function* $u(x, y)$ *such that the inequality*

$$\int_{\lambda_2} R(\sigma)u^2(\sigma)\mathrm{d}\sigma \leq K \int_{\lambda_2} \left(\frac{\partial u}{\partial n}\right)^2 \mathrm{d}\sigma \tag{1}$$

is valid.

We map Ω conformally onto a half plane with the aid of the function $z=\chi(t)$, $t=\xi+i\eta$. We carry out the mapping in such a way that the point $t=\infty$ corresponds to some point within λ_1. The points $\alpha_1, \alpha_2, \ldots, \alpha_{2m}$ then transform into the points $a_1, a_2, \ldots, a_{2m}$ on the real axis ξ. We denote the transforms of $\gamma_k, \lambda_1, \lambda_2$ by Γ_k, Λ_1, Λ_2. We further suppose that $u(x, y)=U(\xi, \eta)$. We denote the function conjugate to $U(\xi, \eta)$ by $V(\xi, \eta)$. Obviously the boundary conditions for $U(\xi, \eta)$ are the following

$$\text{on } \Lambda_1: U(\xi, \eta)=0; \tag{2}$$

$$\text{on } \Lambda_2: \frac{\partial U}{\partial \eta} = \frac{\partial u}{\partial n}|\chi'(\xi)|. \tag{3}$$

It follows from the Cauchy–Riemann equations that on Γ_{2k}

$$V = -\int_{a_{2k}}^{\xi} \frac{\partial U}{\partial \eta}\, \mathrm{d}\xi + C_k = \psi^*(\xi) + C_k; \tag{4}$$

the constants C_k are defined by the continuity conditions associated with the function $U(\xi, \eta)$.

One of the constants C_k can be fixed arbitrarily since $V(\xi, \eta)$ is defined to within an arbitrary constant. We do this by stipulating that C equal zero in the general Schwartz formula for the half plane.

$$U + iV = \frac{1}{\pi i}\int_{-\infty}^{+\infty} \frac{U(\xi, 0)}{\xi - t}\, \mathrm{d}\xi + iC$$

(see, e. g. the author's book [7], § 67, formula (2)). Thus the arbitrariness in the selection of the C_k is entirely eliminated.

Boundary conditions (2) and (4) allow us to find the value of $U(\xi, 0)$ on Λ_2, i.e. on Λ_2

$$U(\xi, 0)=\frac{1}{R_1(\xi)}\left\{\frac{1}{\pi}\int_{\Lambda_2} R_1(\xi')B(\xi')\frac{d\xi'}{\xi'-\xi}+Q_{m-1}(\xi)\right\} \qquad (5)$$

(cf. e.g. the author's book [7], § 67, formula (2)).

Here $R_1(\xi)=\sqrt{\left(\prod_{k=1}^{2m}|\xi-a_k|\right)}$; $Q_{m-1}(\xi)$ is a polynomial of the $(m-1)$th degree:

$$Q_{m-1}(\xi)=q_0\xi^{m-1}+q_1\xi^{m-2}+\ldots+q_{m-1}$$

and, finally,

$$B(\xi)=-\varphi^*(\xi)-C_k.$$

Formula (5) contains $2m$ constants C_k and q_k. We shall prove that after choosing them $U(\xi, \eta)$ is continuous at the points a_k. For the continuity of $U(\xi, \eta)$ it suffices that at $\xi=a_k$ the expression in brackets in formula (5) should vanish. This gives a system of equations in $2m$ unknowns

$$\frac{1}{\pi}\int_{\Lambda_2}\frac{R_1(\xi')}{\xi'-a_k}B(\xi')d\xi'+Q_{m-1}(\alpha_k)=0, \quad k=1, 2, \ldots, 2m. \qquad (6)$$

System (6) has a unique solution; this follows from the uniqueness of the solution of the mixed problem. Solving system (6) we obviously arrive at a solution in the form

$$C_k=\int_{\Lambda_2}\frac{\mu_k(\xi')}{R_1(\xi')}\varphi^*(\xi')d\xi', \quad q_k=\int_{\Lambda_2}\frac{\nu_k(\xi')}{R_1(\xi')}\varphi^*(\xi')d\xi', \qquad (7)$$

where $\mu_k(\xi')$ and $\nu_k(\xi')$ are continuous functions on Λ_2. We define the continuous functions $\tilde{\mu}_k(\xi)$ and $\tilde{\nu}_k(\xi)$ on Λ_2 by putting

$$\tilde{\mu}_k(\xi)=\mu\int_{\xi}^{a_{2s+1}}\frac{\mu_k(\xi')}{R_1(\xi')}d\xi'; \quad \tilde{\nu}_k(\xi)=\int_{\xi}^{a_{2s+1}}\frac{\nu_k(\xi')}{R_1(\xi')}d\xi'$$

on Γ_2.

Integrating (7) by parts and noting that $\varphi^*(a_{2s}) = \tilde{\mu}_k(a_{2s+1}) = \tilde{\nu}_k(a_{2s+1}) = 0$, we have:

$$C_k = \int_{\Lambda_2} \tilde{\mu}_k(\xi) \frac{\partial U}{\partial \eta} \, d\xi, \quad q_k = \int_{\Lambda_2} \tilde{\nu}_k(\xi) \frac{\partial U}{\partial \eta} d\xi.$$

Substituting this in (5) and changing the order of integration in the double integral so obtained, we readily reduce the formula mentioned to the form:

$$U(\xi, 0) = \frac{1}{R_1(\xi)} \int_{\Lambda_2} M(\xi, \xi') \frac{\partial U}{\partial \eta} \, d\xi, \quad (\text{on } \Lambda_2) \tag{8}$$

where $M(\xi, \xi')$ is such that the integral

$$\int_{\Lambda_2} M^2(\xi, \xi') d\xi'$$

is bounded. The latter follows from the fact that the functions $\tilde{\mu}_k(\xi)$, $\tilde{\nu}_k(\xi)$ satisfy a Lipshitz condition with an exponent of $1/2$. $M(\xi, \xi')$ satisfies the same condition within Λ_2 and, *a fortiori*, is logarithmically infinite at the points a_k. (cf. e.g. N. I. Muskhelishvili [1].)

We now apply Cauchy's inequality to (8):

$$\begin{aligned} U^2(\xi, 0) &\leq \frac{2}{R_1^2(\xi)} \int_{\Lambda_2} M^2(\xi, \xi') d\xi' \int_{\Lambda_2} \left(\frac{\partial U}{\partial \eta} \right)^2 d\xi' \\ &< \frac{C'}{R_1^2(\xi)} \int_{\Lambda_2} \left(\frac{\partial U}{\partial \eta} \right)^2 d\xi' \quad (\text{on } \Lambda_2), \; C' = \text{const.} \end{aligned} \tag{9}$$

We now return to the plane $z = x + iy$. The transforming function $z = \chi(t)$ remains finite on Λ_2; then as is readily seen, $\chi'(t)$ and $\frac{1}{\chi'(t)}$ are bounded on Λ_2. But then the ratios $\frac{R_1(\xi)}{R(\sigma)}$ and $\frac{R(\sigma)}{R_1(\xi)}$ are also bounded and we readily find from (9) that on λ_2

$$R(\sigma)u^2(\sigma) < \frac{K''}{R(\sigma)} \int_{\lambda_2} \left(\frac{\partial u}{\partial n}\right)^2 \mathrm{d}\sigma, \quad K'' = \text{const.} \tag{10}$$

Finally, integrating (10) over λ_2 and putting

$$K = K'' \int_{\lambda_2} \frac{\mathrm{d}\sigma}{R(\sigma)},$$

we arrive at inequality (1)

THEOREM 2. *For any values of* u *and* $\frac{\partial u}{\partial n}$ *specified on* λ_1 *and* λ_2 *there exists a constant* K_1 *such that*

$$\int_{\lambda_2} R(\sigma)u^2(\sigma)\mathrm{d}\sigma \leq K_1 \left\{ \int_{\lambda_2} \left(\frac{\partial u}{\partial n}\right)^2 \mathrm{d}\sigma + \int_{\lambda_1} \left[u^2(\sigma) + \left(\frac{\partial u}{\partial \sigma}\right)^2 \right] \mathrm{d}\sigma \right\}. \tag{11}$$

We construct a function $u_1(\sigma)$ which is continuous on S and coincides with $u(\sigma)$ on λ_1; we extend it to λ_2 by merely subjecting it to the two requirements

$$\int_{\lambda_2} u_1^2(\sigma)\mathrm{d}\sigma \leq k \int_{\lambda_1} u^2(\sigma)\mathrm{d}\sigma, \quad \int_{\lambda_2} \left(\frac{\partial u_1}{\partial \sigma}\right)^2 \mathrm{d}\sigma \leq k \int_{\lambda_1} \left(\frac{\partial u}{\partial \sigma}\right)^2 \mathrm{d}\sigma, \tag{12}$$

$k = \text{const.}$

Further we construct a function $u(x, y)$ which is harmonic in Ω and assumes the value $u_1(\sigma)$ on S, and we put $u = u_1 + u_2$. The function u_2 satisfies the condition of theorem 1; by inequality (1):

$$\int_{\lambda_2} R(\sigma)u_2^2(\sigma)\mathrm{d}\sigma \leq K \int_{\lambda_2} \left(\frac{\partial u}{\partial n} - \frac{\partial u_1}{\partial n}\right)^2 \mathrm{d}\sigma \leq 2K \int_{\lambda_2} \left[\left(\frac{\partial u}{\partial n}\right)^2 + \left(\frac{\partial u_1}{\partial n}\right)^2 \right] \mathrm{d}\sigma. \tag{13}$$

We now estimate the last integral in (13). We conformally map Ω onto a circle with the aid of the function $z = \omega(t)$, $t = \varrho e^{i\vartheta}$, $0 \leq \varrho \leq 1$ and we suppose that $u_1(x, y) = U_1(\varrho, \vartheta) = \mathrm{Re}\{F(t)\}$.

We expand $F(t)$ into a power series

$$F(t) = \sum_{n=0}^{\infty} a_n t^n = \sum_{n=0}^{\infty} a_n \varrho^n e^{in\vartheta}.$$

Then

$$U_1(\varrho, \vartheta)=\frac{1}{2}\sum_{n=0}^{\infty}\varrho^n(a_n e^{in\vartheta}+\bar{a}_n e^{-in\vartheta}).$$

Hence

$$\left.\frac{\partial U_1(\varrho, \vartheta)}{\partial \varrho}\right|_{\varrho=1}=\frac{1}{2}\sum_{n=1}^{\infty}n(a_n e^{in\vartheta}+\bar{a}_n e^{-in\vartheta}),$$

$$\left.\frac{\partial U_1(\varrho, \vartheta)}{\partial \vartheta}\right|_{\varrho=1}=\frac{i}{2}\sum_{n=1}^{\infty}n(a_n e^{in\vartheta}-\bar{a}_n e^{-in\vartheta})$$

and, consequently, by virtue of Parseval's equation,

$$\int_{\Gamma}\left[\left.\frac{\partial U_1(\varrho, \vartheta)}{\partial \varrho}\right|_{\varrho=1}\right]^2 d\vartheta=\int_{\Gamma}\left[\left.\frac{\partial U_1(\varrho, \vartheta)}{\partial \vartheta}\right|_{\varrho=1}\right]^2 d\vartheta$$

or, reverting to the old variables:

$$\int_S\left(\frac{\partial u_1}{\partial n}\right)^2|\omega'(\tau)|\,d\sigma=\int_S\left(\frac{\partial u_1}{\partial \sigma}\right)^2|\omega'(\tau)|\,d\sigma.$$

Let $A=\min|\omega'(t)|$, $B=\max|\omega'(t)|$. Obviously $0<A\leq B<\infty$ since the contour S is smooth. Then from the last equation it follows that

$$\int_S\left(\frac{\partial u_1}{\partial n}\right)^2 d\sigma\leq\frac{B}{A}\int_S\left(\frac{\partial u_1}{\partial \sigma}\right)^2 d\sigma.$$

A fortiori

$$\int_{\lambda_2}\left(\frac{\partial u_1}{\partial n}\right)^2 d\sigma\leq\frac{B}{A}\left\{\int_{\lambda_1}\left(\frac{\partial u_1}{\partial \sigma}\right)^2 d\sigma+\int_{\lambda_2}\left(\frac{\partial u_1}{\partial \sigma}\right)^2 d\sigma\right\}$$

and, by virtue of the definition of $u_1(\sigma)$ and of inequalities (12)

$$\int_{\lambda_2}\left(\frac{\partial u_1}{\partial n}\right)^2 d\sigma\leq\frac{B(1+k)}{A}\int_{\lambda_1}\left(\frac{\partial u}{\partial \sigma}\right)^2 d\sigma.$$

Substituting this in (13) we obtain:

$$\int_{\lambda_2} R(\sigma)u_2^2(\sigma)\mathrm{d}\sigma \leq 2K\int_{\lambda_2}\left(\frac{\partial u}{\partial n}\right)^2\mathrm{d}\sigma + K'\int_{\lambda_1}\left(\frac{\partial u}{\partial \sigma}\right)^2\mathrm{d}\sigma,$$

$$K' = \frac{2BK(1+k)}{A}. \tag{14}$$

Now the inequality (11) is readily proved. We have:

$$\int_{\lambda_2} R(\sigma)u^2(\sigma)\mathrm{d}\sigma = \int_{\lambda_2} R(\sigma)[u_1(\sigma)+u_2(\sigma)]^2\mathrm{d}\sigma$$

$$\leq 2R^*\int_{\lambda_2} u_1^2(\sigma)\mathrm{d}\sigma + 2\int_{\lambda_2} R(\sigma)u_2^2(\sigma)\mathrm{d}\sigma,$$

where $R^* = \max R(\sigma)$. We use inequalities (12) and (14) and put $K_1 = \max(2kR^*, 4K, 2K')$. We then obtain:

$$\int_{\lambda_2} R(\sigma)u^2(\sigma)\,\mathrm{d}\sigma \leq K_1\left\{\int_{\lambda_2}\left(\frac{\partial u}{\partial n}\right)^2\mathrm{d}\sigma + \int_{\lambda_1}\left[\left(\frac{\partial u}{\partial \sigma}\right)^2 + u^2(\sigma)\right]\mathrm{d}\sigma\right\}.$$

We now pass on to the solution of the mixed problem of potential theory. We suppose that $v(x, y)$ is the function conjugate to the $u(x, y)$ required which vanishes at the coordinate origin, and that $f(z) = u + iv$.

We shall look for an approximate solution in the form

$$f_n(z) = u_n + iv_n = \sum_{k=0}^{n} a_k z^k, \quad \mathrm{Im}(a_0) = 0.$$

We determine the coefficients of the polynomial $f_n(z)$ from the condition that

$$\int_{\lambda_1}\left[(u-u_n)^2 + \left(\frac{\partial u}{\partial \sigma} - \frac{\partial u_n}{\partial \sigma}\right)^2\right]\mathrm{d}\sigma + \int_{\lambda_2}\left(\frac{\partial u}{\partial n} - \frac{\partial u_n}{\partial n}\right)^2\mathrm{d}\sigma \tag{15}$$

is a minimum. Conditions (15) lead to a system of linear equations in the unknowns a_k. We shall prove that this system is soluble. We consider the Hilbert space whose elements are functions which are harmonic in Ω and representable in

terms of Green's function in which the scalar product is defined by the formula

$$(u, v)=\int_{\lambda_1}\left(u\bar{v}+\frac{\partial u}{\partial\sigma}\frac{\partial\bar{v}}{\partial\sigma}\right)d\sigma+\int_{\lambda_2}\frac{\partial u}{\partial n}\frac{\partial\bar{v}}{\partial n}d\sigma.$$

Such a definition is satisfactory. In fact, the verification only requires axiom D (§ 3). But if $(u, u)=0$ then $u\equiv 0$ on λ_1 and $\frac{\partial u}{\partial n}\equiv 0$ on λ_2. By the theorem on the uniqueness of the solution of the mixed problem of elasticity theory, $u\equiv 0$.

Now the minimal condition (15) assumes the form $\|u-u_n\|^2=\min$. Since u_n is a linear combination of the linearly independent functions 1, $\mathrm{Re}(z^k)$, $\mathrm{Im}(z^k)$, then as proved in § 83, the linear system to which condition (15) reduces is uniquely soluble.

We now prove that $f_n(z)\to f(z)$ *uniformly within* Ω. Firstly, using corollary 2 of theorem 1 § 86 we shall prove, as in the previous section, that each of the integrals in (15) tends to zero and then, by virtue of inequality (11), the integral

$$\int_{\lambda_2}R(\sigma)(u-u_n)^2d\sigma$$

also tends to zero. Now

$$\int_S R(\sigma)(u-u_n)^2d\sigma\leq R^*\int_{\lambda_1}(u-u_n)^2d\sigma+\int_{\lambda_2}R(\sigma)(u-u_n)^2d\sigma\to 0.$$

Further

$$f(z)-f_n(z)=\frac{1}{2\pi}\int_S(u-u_n)T(z;\zeta)d\sigma$$

$$=\frac{1}{2\pi}\int_S(\sqrt{R(\sigma)})(u-u_n)\frac{T(z;\zeta)}{\sqrt{R(\sigma)}}d\sigma,$$

where $T(z;\zeta)$ is Schwartz's kernel for the domain Ω. (For the definition and properties of Schwartz's kernel see the author's book [7].) By Cauchy's inequality

$$|f(z)-f_n(z)|^2\leq\frac{1}{4\pi^2}\int_S\frac{T^2(z;\zeta)}{R(\sigma)}d\sigma\int_S R(\sigma)(u-u_n)^2d\sigma. \qquad (16)$$

If $z \in \bar{\Omega}'$, where $\bar{\Omega}'$ is a closed domain lying within Ω, then $T(z, \zeta)$ is continuous, the first integral in (16) is bounded and since the second integral tends to zero, $f_n(z) \to f(z)$ uniformly within Ω.

§ 91. The plane problem of elasticity theory

We shall restrict ourselves to the case of specified displacements; the case of specified stresses is not essentially different. As before we shall restrict ourselves here to a singly-connected finite domain.

As is known (cf. N. I. Muskhelishvili [2]) our problem reduces to the seeking of two functions $\varphi(z)$ and $\psi(z)$ which are holomorphic in the considered domain Ω and are related on the contour S of the domain Ω by the relation

$$\varkappa\varphi(\zeta) - \zeta\overline{\varphi'(\zeta)} - \overline{\psi(\zeta)} = g(\zeta). \tag{1}$$

Here $\varkappa = 3 - 4\sigma$, where σ is Poisson's ratio, $g(\zeta)$ is a specified function which is continuous on S. The solution of the boundary value problem thus formulated exists (cf. N. I. Muskhelishvili [2]). We can also stipulate that $\varphi(0) = 0$, $\operatorname{Im}\{\varphi(0)\} = 0$. It can be proved (cf. the author's paper [1]; the proof given in this paper can be simplified) that the value of $\varphi(z)$ can be expressed in terms of $g(z)$ on the contour by a formula of the form

$$\varphi(z) = \frac{1}{2\varkappa} g(z) + \frac{1}{2\pi i \varkappa} \int_S \frac{g(\zeta)}{\zeta - z} d\zeta + \int_S \{K_1(z, \zeta) g(\zeta) + K_2(z, \zeta)\overline{g(\zeta)}\} d\zeta, \tag{2}$$

where K_1 and K_2 are continuous functions of the points z and ζ on the curve S. By virtue of lemma 1 § 86, there readily follows the estimate

$$\|\varphi\| \leq C\|g\|, \quad C = \text{const}. \tag{3}$$

We now replace all the terms in (1) by the complex conjugates and multiply by

$$\frac{1}{2\pi i}\frac{d\zeta}{\zeta-z}, \quad z\in\Omega$$

and integrate over S. This gives

$$\psi(z)=-\frac{1}{2\pi i}\int_S\frac{\overline{g(\zeta)}}{\zeta-z}d\zeta-\frac{1}{2\pi i}\int_S\frac{\overline{\zeta}\varphi'(\zeta)}{\zeta-z}d\zeta+\frac{\varkappa}{2\pi i}\int_S\frac{\overline{\varphi(\zeta)}}{\zeta-z}d\zeta,$$

or, integrating the second integral by parts,

$$\psi(z)=-\frac{1}{2\pi i}\int_S\frac{\overline{g(\zeta)}}{\zeta-z}d\zeta+\frac{1}{2\pi i}\int_S\varphi(\zeta)d\left(\frac{\overline{\zeta}}{\zeta-z}\right)+\frac{\varkappa}{2\pi i}\int_S\frac{\overline{\varphi(\zeta)}}{\zeta-z}d\zeta.$$

If z lies within a closed domain $\overline{\Omega}'$ which lies entirely within Ω, then, applying Cauchy's inequality to the last formula we readily find

$$|\psi(z)|\leq C'\|g\|, \quad C'=\text{const.} \tag{4}$$

We shall seek approximations to the functions $\varphi(z)$ and $\psi(z)$ in the form of the polynomials

$$\varphi_n(z)=\sum_{k=1}^{n}\alpha_k z^k, \quad \psi_n(z)=\sum_{k=0}^{n}\beta_k z^k, \tag{5}$$

whose coefficients we find from the condition

$$\|g_n-g\|^2=\|\varkappa\varphi_n(z)-z\overline{\varphi_n'(z)}-\overline{\psi_n(z)}-g\|^2=\min. \tag{6}$$

As usual, this condition leads to a system of linear algebraic equations. We shall prove that this system is soluble uniquely.

We shall change the notation for the polynomials $\varphi_n(z)$ and $\psi_n(z)$ and write

$$\varphi_n(z)=\sum_{k=1}^{n}(a_{2k-1}+ia_{2k})z^k,$$

$$\psi_n(z)=\sum_{k=0}^{n}(a_{2k+2n+1}+ia_{2k+2n+2})z^k,$$

where this time the a_j denote real quantities. We further introduce the notation:

$$\varphi_{(2k-1)}(z)=z^k, \quad 1\leq k\leq n; \quad \varphi_{(2k)}(z)=iz^k, \quad 2\leq k\leq n;$$
$$\varphi_{(k)}(z)=0, \quad k<2n; \quad \psi_{(k)}(z)=0, \quad 1\leq k\leq 2n;$$
$$\psi_{(2k+2n+1)}(z)=z^k, \quad 0\leq k\leq n; \quad \psi_{(2k+2n+2)}(z)=iz^k, \quad 0\leq k\leq n.$$

Then, obviously,

$$\varphi_n(z)=\sum_{k=1}^{4n+2} a_k\varphi_{(k)}(z), \quad \psi_n(z)=\sum_{k=1}^{4n+2} a_k\psi_{(k)}(z),$$

$$g_n(z)=\sum_{k=1}^{4n+2} a_k g_{(k)}(z),$$

where

$$g_{(k)}(z)=\varkappa\varphi_{(k)}(z)-\overline{z\varphi'_{(k)}(z)}-\overline{\psi_{(k)}(z)}.$$

The identity

$$\sum_{k=1}^{4n+2} b_k g_{(k)}(\zeta)\equiv 0, \quad \zeta\in S,$$

where b_k are real numbers, is only possible if all these numbers equal zero.

In fact if

$$\sum_{k=1}^{4n+2} b_k g_{(k)}(\zeta)\equiv 0,$$

then simultaneously, by the theorem on the uniqueness of the solution of the elasticity theory problem,

$$\sum_{k=1}^{4n+2} b_k\varphi_{(k)}(z)\equiv 0, \quad \sum_{k=1}^{4n+2} b_k\psi_{(k)}(z)\equiv 0,$$

or

$$\sum_{k=1}^{n}(b_{2k-1}+ib_{2k})z^k\equiv 0, \quad \sum_{k=1}^{n}(b_{2n+2k+1}+ib_{2k+2n+2})z^k\equiv 0,$$

whence it follows that all the b_j equal zero.

Condition (6) now assumes the form

$$\left\|\sum_{k=1}^{4n+2} a_k g_{(k)}-g\right\|^2=\min;$$

or, since the numbers a_k are real,

$$\sum_{j,k=1}^{4n+2} a_j a_k (g_{(j)}, g_{(k)}) - 2\sum_{k=1}^{4n+2} a_k \mathrm{Re}(g, g_{(k)}) + (g, g) = \min.$$

Differentiating this with respect to a_j and equating the result to zero, we obtain a system of linear equations in the unknowns a_k:

$$\sum_{k=1}^{4n+2} a_k \mathrm{Re}(g_{(k)}, g_{(j)}) = \mathrm{Re}(g, g_{(j)}), \quad j=1, 2, \ldots, 4n+2. \quad (7)$$

We shall prove that system (7) is soluble, for this purpose we consider the homogeneous system

$$\sum_{k=1}^{4n+2} b_k \mathrm{Re}(g_{(k)}, g_{(j)}) = 0, \quad j=1, 2, \ldots, 4n+2. \quad (8)$$

Putting

$$\tilde{g}(\zeta) = \sum_{k=1}^{4n+2} b_{(k)} g_{(k)}(\zeta),$$

it is possible to write it in the form

$$\mathrm{Re}(\tilde{g}, g_{(j)}) = 0, \quad j=1, 2, \ldots, 4n+2.$$

Multiplying by b_j and adding we obtain, recalling that the b_k are real:

$$\mathrm{Re}(\tilde{g}, \tilde{g}) = 0.$$

But $\mathrm{Re}(\tilde{g}, \tilde{g}) = \mathrm{Re}\|\tilde{g}\|^2 = \|\tilde{g}\|^2$. Hence $\tilde{g}=0$ and, as proved above $b_k=0$, $k=1, 2, \ldots, 4n+2$. System (8) thus has only a trivial solution. Hence it follows that the non-homogeneous system (7) is always soluble.

We now study the convergence of the approximate solution. If the given function $g(\zeta)$ is sufficiently smooth (it suffices that $g_1(\zeta)$ satisfy a Lipshitz condition with a positive exponent) then $\varphi'(z)$ and $\psi(z)$ will be continuous in Ω. By using the corollaries of Walsh's theorem and the theorem on the existence and uniqueness of our problem we readily find that $\|g_n - g\| \to 0$. On the basis of inequality (3), $\|\varphi_n - \varphi\| \to 0$ and by virtue of theorem 2 § 86, $\varphi_n(z) \to \varphi(z)$ uniformly within

Ω. Finally it follows from (4) that $\psi_n(z) \to \psi(z)$ uniformly within Ω.

We shall have uniform convergence in $\bar{\Omega}$ if we require that the quantity $\|g_n - g\|^2$ shall be a minimum.

The results of the present section can be extended to the case of an anisotropic medium without difficulty.

§ 92. The periodic problem of elasticity theory

We consider the singly-connected domain Ω of the plane $z = x + iy$ which is bounded by the curve $y = y(x)$ where $y(x)$ is a continuous periodic function of x which is differentiable a sufficient number of times. We denote the period of $y(x)$ by L. Suppose that an external force which also varies periodically with the same period L is applied to the boundary of the domain. As has been established (cf. the author's book [7], § 55), the functions $\varphi(z)$ and $\psi(z)$ can be represented in the form

$$\varphi(z) = \varphi_0(z) + \beta z, \quad \psi(z) = \psi_0(z) - z\varphi'(z) + \bar{\beta}\varkappa z. \tag{1}$$

Here $\varphi_0(z)$ and $\psi_0(z)$ are functions which are periodic with period L, $\varkappa = 3 - 4\sigma$ (σ is Poisson's ratio) under the conditions of plane deformation and $\varkappa = \dfrac{3-\sigma}{1+\sigma}$ under the conditions of the generalized plane strained state. Finally

$$\beta = \frac{i(X + iY)}{(1+\varkappa)L}, \tag{2}$$

where X and Y are the components along the x and y axes of the principal external force vector which is applied from the side of the positive normal to an arc of the contour corresponding to one period (we shall denote this arc by C).

By substituting (1) and (2) in the boundary condition for the plane problem of elasticity theory

$$\varphi(\zeta) + \zeta\overline{\varphi'(\zeta)} + \overline{\psi(\zeta)} = i\int_{\zeta_0}^{\zeta} (X_n + iY_n)\,d\sigma, \quad d\sigma = |d\zeta|,$$

where ζ is a point on the arc C, X_n and Y_n the components of the stress acting upon this arc, we reduce this condition to the form

$$\varphi_0(\zeta)+(\zeta-\overline{\zeta})\,\overline{\varphi_0'(\zeta)}+\overline{\psi_0(\zeta)}=f(\zeta), \tag{3}$$

where $f(\zeta)$ is some specified continuous periodic function of ζ.

The transformation

$$e^{\frac{2\pi i}{L}z}=t \tag{4}$$

transforms the strip $0\leq x\leq L$ into the t-plane cut along the positive real semi-axis, and the arc C into some closed contour Γ. We shall assume that the elastic medium occupies the region above the curve $y=y(x)$. The part of the pole $0\leq x\leq L$ which is filled by the elastic medium then transforms into the domain Σ' which is located within Γ and cut along the positive real axis. The functions $\Phi(t)=\varphi_0\left(\frac{L}{2\pi i}\ln t\right)$ and $\Psi(t)=\psi_0\left(\frac{L}{2\pi i}\ln t\right)$ are holomorphic in Σ'. Being periodic in z these functions assume the same values at geometrically equivalent points on the profile and, consequently, are defined everywhere along the profile. But then, as is known, these functions can be continued analytically across the profile and are holomorphic in the entire domain Σ included within Γ. The problem reduces to the determination of the functions $\Phi(t)$ and $\Psi(t)$ which are holomorphic in Σ and satisfy the boundary condition

$$\Phi(\tau)-\frac{\tau}{\overline{\tau}}\ln|\tau|\,\overline{\Phi'(\tau)}+\overline{\Psi(\tau)}=F(\tau);\quad F(\tau)=f\left(\frac{L}{2\pi i}\ln\tau\right) \tag{5}$$

on Γ. We arrive at equation (5) if in (3) we put

$$\zeta=\frac{L}{2\pi i}\ln\tau.$$

We apply to our problem a procedure similar to that which D. I. Sherman [1] (cf. also the author's book [7] § 56) used in

his derivation of Lauricella's equations. We shall seek $\Phi(t)$ and $\Psi(t)$ in the form

$$\Phi(t)=\frac{1}{2\pi i}\int_\Gamma \frac{\omega(\sigma)}{\sigma-t}\,d\sigma, \quad \Phi'(t)=\frac{1}{2\pi i}\int_\Gamma \frac{\omega'(\sigma)}{\sigma-t}\,d\sigma.$$

$$\Psi(t)=\frac{1}{2\pi i}\int_\Gamma \frac{\overline{\omega(\sigma)}}{\sigma-t}\,d\sigma+\frac{1}{\pi i}\int_\Gamma \frac{\sigma\ln|\sigma|}{\sigma-t}\,\omega'(\sigma)d\sigma, \tag{6}$$

where σ denotes the complex coordinate of a point on Γ and $\omega(\sigma)$ is an unknown function of σ.

We now construct the expression

$$\Phi(t)-2\bar{t}\ln|t|\overline{\Phi'(t)}+\overline{\Psi(t)},$$

where t is a point within Γ:

$$\Phi(t)-2\bar{t}\,\ln|t|\overline{\Phi'(t)}+\overline{\Psi(t)}=\frac{1}{2\pi i}\int_\Gamma \frac{\omega(\sigma)}{\sigma-t}\,d\sigma$$

$$+\frac{\bar{t}\ln|t|}{\pi i}\int_\Gamma \frac{\overline{\omega'(\sigma)}}{\bar{\sigma}-\bar{t}}\,d\bar{\sigma}-\frac{1}{2\pi i}\int_\Gamma \frac{\omega(\sigma)}{\bar{\sigma}-\bar{t}}\,d\bar{\sigma}-\frac{1}{\pi i}\int_\Gamma \frac{\bar{\sigma}\ln|\sigma|}{\bar{\sigma}-\bar{t}}\,\overline{\omega'(\sigma)}\,d\bar{\sigma}.$$

We now suppose that $t\to\tau$ where τ is a point on the contour of Γ. Using the well-known theorem on the limit of a Cauchy-type integral, we obtain

$$\Phi(\tau)-2\bar{\tau}\ln|\tau|\overline{\Phi'(\tau)}+\overline{\Psi(\tau)}$$

$$=\omega(\tau)+\frac{1}{2\pi i}\int_\Gamma \omega(\sigma)\,d\ln\frac{\sigma-\tau}{\bar{\sigma}-\bar{\tau}}-\frac{1}{\pi i}\int_\Gamma \frac{\overline{\bar{\sigma}\ln|\sigma|-\bar{\tau}\ln|\tau|}}{\bar{\sigma}-\bar{\tau}}\,d\overline{\omega(\sigma)}.$$

By virtue of equation (3) the expression obtained is equal to $F(\tau)$. Now we put $\sigma-\tau=re^{i\vartheta}$ in the first integral and integrate the second by parts. We then obtain an integral equation equivalent to a system of two Fredholm-type equations with continuous kernels

$$\omega(\tau)+\frac{1}{\pi}\int_\Gamma \omega(\sigma)\,d\vartheta+\frac{1}{\pi i}\int_\Gamma \overline{\omega(\sigma)}\,\frac{d}{d\bar{\sigma}}\,\frac{\bar{\sigma}\ln|\sigma|-\bar{\tau}\ln|\tau|}{\bar{\sigma}-\bar{\tau}}\,d\bar{\sigma}=F(\tau). \tag{7}$$

(We obtain this system by separating the real and imaginary parts in (7) and taking $\mathrm{Re}\{\omega(\tau)\}$ and $\mathrm{Im}\{\omega(\tau)\}$ as the unknowns.)

We shall prove that equation (7) is soluble for any function $F(\tau)$.

For this purpose, we consider the homogeneous equation

$$\omega_0(\tau)+\frac{1}{\pi}\int_\Gamma \omega_0(\sigma)\,d\vartheta+\frac{1}{\pi i}\int_\Gamma \overline{\omega_0(\sigma)}\,\frac{d}{d\bar\sigma}\,\frac{\bar\sigma\ln|\sigma|-\bar\tau\ln|\tau|}{\bar\sigma-\bar\tau}\,d\bar\sigma=0, \quad (8)$$

and suppose that it has some solution $\omega_0(\sigma)$. We put

$$\Phi_0(t)=\frac{1}{2\pi i}\int_\Gamma \frac{\omega_0(\sigma)}{\sigma-t}\,d\sigma,$$

$$\Psi_0(t)=\frac{1}{2\pi i}\int_\Gamma \frac{\overline{\omega_0(\sigma)}}{\sigma-t}\,d\sigma+\frac{1}{\pi i}\int_\Gamma \frac{\sigma\ln|\sigma|}{\sigma-t}\,\omega_0'(\sigma)\,d\sigma. \quad (9)$$

These functions solve the homogeneous problem of elasticity theory, therefore if we put

$$\varphi_0(z)=\Phi\left(e^{\frac{2\pi i}{L}z}\right), \quad \psi_0(z)=\Psi\left(e^{\frac{2\pi i}{L}z}\right),$$

$$\psi(z)=\psi_0(z)-z\varphi_0'(z),$$

then necessarily

$$\varphi_0(z)=iaz+b, \quad \psi(z)=-\bar b$$

and consequently $\psi_0(z)=-iaz-\bar b$, where a and b are constants. Since $\varphi_0(z)$ and $\psi_0(z)$ are periodic, $a=0$ and thus

$$\varphi_0(z)=b, \quad \psi_0(z)=-\bar b.$$

Substituting this in (9), we obtain the identities

$$\frac{1}{2\pi i}\int_\Gamma \frac{\omega_0(\sigma)-b}{\sigma-t}\,d\sigma=0,$$

$$\frac{1}{2\pi i}\int_\Gamma \frac{\overline{\omega_0(\sigma)}+2\sigma\ln|\sigma|\omega_0'(\sigma)+b}{\sigma-t}\,d\sigma=0, \quad (10)$$

which are valid if t lies within Γ. Hence it follows that the functions $\delta(t)$ and $\varepsilon(t)$ defined on Γ by the equations

$$\begin{aligned} i\delta(\sigma) &= \omega_0(\sigma) - \mathrm{b}, \\ i\varepsilon(\sigma) &= \overline{\omega_0(\sigma)} + 2\sigma \ln|\sigma|\omega_0'(\sigma) + \overline{b}, \end{aligned} \tag{11}$$

are holomorphic outside Γ and vanish at $t=\infty$. Eliminating $\omega_0(\tau)$ from (11), we obtain

$$\delta(\sigma) - 2\overline{\sigma}\ln|\sigma|\,\overline{\delta'(\sigma)} + \overline{\varepsilon(\sigma)} - 2\overline{b}i = 0. \tag{12}$$

The domain of the t-plane which lies outside Γ corresponds on the z-plane to the part of the strip $0 \leq x \leq L$ lying under C. Equation (12) shows that $\delta(t)$ and $\varepsilon(t) + 2\overline{b}i$ solve the periodic problem of elasticity theory in the absence of external forces on the domain which is underneath the curve $y \equiv y(x)$. But then

$$\delta(t) = b', \quad \varepsilon(t) + 2\overline{b}i = -\overline{b}'.$$

Since $\delta(\infty) = 0$, $b' = 0$ and, consequently, $\delta(\sigma) = 0$. Further, $\varepsilon(\infty) = 0$ and thus $b = 0$; now from (11), it follows that $\omega_0(\sigma) \equiv 0$. The homogeneous integral equation (8) has the unique solution $\omega_0(\sigma) \equiv 0$. Hence it follows that equation (7) has a solution which is unique no matter what its right-hand side.

Separating the real and imaginary parts in (7) and taking $\omega_1(\tau) = \mathrm{Re}\{\omega(\tau)\}$ and $\omega_2(\tau) = \mathrm{Im}\{\omega(\tau)\}$ as the unknown functions we obtain a system of Fredholm equations of the form

$$\begin{aligned} \omega_1(\tau) + \int_\Gamma \{K_{11}(\tau, \sigma)\omega_1(\sigma) + K_{12}(\tau, \sigma)\omega_2(\sigma)\}\,\mathrm{d}s_\sigma &= F_1(\tau), \\ \omega_2(\tau) + \int_\Gamma \{K_{21}(\tau, \sigma)\omega_1(\sigma) + K_{22}(\tau, \sigma)\omega_2(\sigma)\}\mathrm{d}s_\sigma &= F_2(\tau), \end{aligned} \tag{13}$$

$$\mathrm{d}s^\sigma = |\mathrm{d}\sigma|,$$

where the kernels $K_{mn}(\tau, \sigma)$ are continuous and we have used the notation

$$F_1(\tau) = \mathrm{Re}\{F(\tau)\}, \quad F_2(\tau) = \mathrm{Im}\{F(\tau)\}.$$

As we have proved, system (13) is soluble uniquely. The resolvent kernels of this system which we denote by $R_{mn}(\tau, \sigma)$ are continuous by virtue of the continuity of the kernels $K_{mn}(\tau, \sigma)$. The solution of system (13) has the form

$$\omega_1(\tau) = F_1(\tau) + \int_\Gamma \{R_{11}(\tau, \sigma) F_1(\sigma) + R_{12}(\tau, \sigma) F_2(\sigma)\} ds_\sigma,$$

$$\omega_2(\tau) = F_2(\tau) + \int_\Gamma \{R_{21}(\tau, \sigma) F_1(\sigma) + R_{22}(\tau, \sigma) F_2(\sigma)\} ds_\sigma.$$

We multiply the second expression by i and add it to the first. Using the notation $R_{11} + iR_{21} = R_1$, $R_{12} + iR_{22} = R_2$, we then have

$$\omega(\tau) = F(\tau) + \int_\Gamma \{R_1(\tau, \sigma) F_1(\sigma) + R_2(\tau, \sigma) F_2(\sigma)\} ds_\sigma. \qquad (14)$$

As the fundamental space of our problem we introduce the Hilbert space $L_2(\Gamma)$. (With equal success we could introduce the space $L_2(p; \Gamma)$ with any weight function p which is bounded above and below by positive numbers. We shall use this in § 93.) The kernels R_1 and R_2 being continuous are bounded.

Hence it follows that

$$\int_\Gamma \int_\Gamma |R_k(\tau, \sigma)|^2 ds_\sigma ds_\tau < B^2, \quad k = 1, 2,$$

where B is some constant. From formula (14) it follows that $\|\omega\| \leq \|F\| + B(\|F_1\| + \|F_2\|)$. Further

$$\|F\|^2 = \int_\Gamma |F(\tau)|^2 ds = \int_\Gamma \{F_1^2(\tau) + F_2^2(\tau)\} ds = \|F_1\|^2 + \|F_2\|^2.$$

Hence $\|F_1\| \leq \|F\|$, $\|F_2\| \leq \|F\|$. This gives

$$\|\omega\| \leq B_1 \|F\|, \quad B_1 = 1 + 2B. \qquad (15)$$

From the second of formulae (6) it follows that in any closed domain which lies entirely within Ω an estimate of the form

$$|\Phi(t)| \leq D_1 \|\omega\|,$$

$$|\Psi(t)| \leq D_1 \|\omega\|, \quad D_1, D_2 = \text{const},$$

is valid (in order to obtain the estimate for $|\Psi(t)|$ the second integral in the expression for $\Psi(t)$ should be integrated by parts), or, using estimate (15),

$$|\Phi(t)| \leq D\|F\|,$$
$$|\Psi(t)| \leq D'\|F\|. \tag{16}$$

Further, again on the basis of formula (15) and of lemma 1 § 86

$$\|\Phi\| \leq D_0\|F\|, \qquad D_0 = \text{const.} \tag{17}$$

We now make the following observation. $\Phi(t)$ and $\Psi(t)$ are not quite uniquely determined from boundary condition (5). It is possible to replace $\Phi(t)$ by $\Phi(t)+A$ and $\Psi(t)$ by $\Psi(t)+A$ where A is an arbitrary constant. In our method of solution this arbitrariness has been eliminated thanks to a particular specification of the functions $\Phi(t)$ and $\Psi(t)$ by means of formulae (6), and estimates (16) and (17) have been established just for functions defined by these formulae. We now prove that the estimates mentioned remain valid (of course with revised values of the constants D, D' and D_0) if $\Phi(t)$ is defined so that $\Phi(0)=0$.

In fact, for this definition, it is sufficient to subtract the constant

$$A = \frac{1}{2\pi i}\int_\Gamma \frac{\omega(\sigma)}{\sigma}\,d\sigma$$

from the function $\Phi(t)$ defined by formula (6). It is necessary to add

$$\overline{A} = -\frac{1}{2\pi i}\int_\Gamma \frac{\overline{\omega(\sigma)}}{\overline{\sigma}}\,d\overline{\sigma}$$

to the function $\Psi(t)$. Denoting the new functions by $\Phi_1(t)$, $\Psi_1(t)$ we have

$$\Phi_1(t) = \frac{1}{2\pi i}\int_\Gamma \frac{\omega(\sigma)}{\sigma - t}\,d\sigma - \frac{1}{2\pi i}\int_\Gamma \frac{\omega(\sigma)}{\sigma}\,d\sigma = \Phi(t) - A,$$

$$\Psi_1(t)=\frac{1}{2\pi i}\int_\Gamma \frac{\overline{\omega(\sigma)}}{\sigma-t}\,d\sigma$$

$$+\frac{1}{\pi i}\int_\Gamma \frac{\sigma\ln(\sigma)}{\sigma-t}\,\omega'(\sigma)\,d\sigma-\frac{1}{2\pi i}\int_\Gamma \frac{\overline{\omega(\sigma)}}{\bar{\sigma}}\,d\bar{\sigma}=\Psi(t)+\bar{A}.$$

Now

$$\|\Phi_1(t)\|\le\|\Phi(t)\|+\|A\|\le D_0\|F\|+\|A\|.$$

Further

$$\|A\|^2=\int_\Gamma |A|^2 ds_\sigma=|A|^2 l, \tag{18}$$

where l is the length of the contour Γ. On the other hand

$$|A|^2\le\frac{1}{(2\pi)^2}\left\{\int_\Gamma \frac{|\omega(\sigma)|}{|\sigma|}\,ds_\sigma\right\}^2\le\frac{1}{4\pi^2}\int_\Gamma \frac{ds_\sigma}{|\sigma|^2}\int_\Gamma |\omega(\sigma)|^2 ds_\sigma.$$

We denote by δ the shortest distance from the coordinate origin to Γ. Then $\frac{1}{\omega(\sigma)}\le\frac{1}{\delta}$ and $|A|^2=\frac{l}{4\pi^2\delta^2}\ \|\omega\|^2$

Substituting this in (18) and using estimate (15), we obtain $\|A\|\le\frac{B_1 l}{2\pi\delta}\|F\|$. Hence

$$\|\Phi_1\|\le C_0\|F\|,\quad C_0=D_0+\frac{B_1 l}{2\pi\delta}.$$

Thus $\Phi_1(t)$ has an estimate of type (17). In precisely the same way it can be proved that $\Phi_1(t)$ and $\Psi_1(t)$ have estimates of type (16).

Estimates (16) and (17) allow us to apply the method of least squares to our problem. We put approximately

$$\Phi(t)\approx\Phi_n(t)=\sum_{k=1}^{n} a_k t^k,\quad \Psi(t)\approx\Psi_n(t)=\sum_{k=0}^{n} b_k t^k.$$

We use the notation

$$F_n(\tau)=\Phi_n(\tau)-2\bar{\tau}\ln|\tau|\overline{\Phi_n'(\tau)}+\overline{\Psi_n(\tau)}$$

and determine the coefficients a_k and b_k from the condition

$$\|F-F_n\|^2=\min. \tag{19}$$

Repeating the reasoning of § 91, we find that $\|F-F_n\|\to 0$. Hence on the basis of estimates (16) and (17), it follows that $\|\Phi-\Phi_n\|\to 0$ and that $\Phi_n(t)\to\Phi(t)$, $\Psi_n(t)\to\Psi(t)$ uniformly in any domain which lies entirely within Γ.

§ 93. Stress in an elastic space bounded by a sine curve

As an illustration of the method we consider the case where the elastic medium occupies the domain above the sinusoidal curve $y=\sin x$.

In order to simplify the calculation we suppose that only normal constraints are applied to the boundary of the domain and that they vary periodically according to the law

$$Y_n=-\frac{q\cos x}{\sqrt{(1+\cos^2 x)}},\quad q=\text{const.} \tag{1}$$

Then

$$f(s)=i\int(X_n+iY_n)\,ds=q\sin x. \tag{2}$$

The transformation

$$t=e^{iz} \tag{3}$$

transforms the arc of the sinusoid $0\le x\le 2\pi$ into the contour Γ whose equation is

$$\tau=e^{i(x+i\sin x)}=e^{-\sin x}(\cos x+i\sin x) \tag{4}$$

(here x plays the part of a parameter).

Equation (5) § 92 in our case assumes the form

$$\Phi(\tau)-2\bar{\tau}\ln|\tau|\overline{\Phi'(\tau)}+\overline{\Psi(\tau)}=-q\ln|\tau|. \tag{5}$$

The contour Γ is symmetric about the imaginary axis, therefore if the point τ lies on Γ, then the point $-\bar{\tau}$ also lies on Γ. We replace τ by $-\bar{\tau}$ in (5) and replace all the terms in the equation so obtained by their conjugates:

$$\overline{\Phi(-\bar{\tau})}+2\bar{\tau}\ln|\tau|\Phi'(-\bar{\tau})+\Psi(-\bar{\tau})=-q\ln|\tau|. \tag{6}$$

Equation (6) shows that our problem is solved by $\overline{\Phi(-t)}$ and $\overline{\Psi(-t)}$ as well as by $\Phi(t)$ and $\Psi(t)$. But the functions $\Phi(t)$ and $\Psi(t)$ are defined uniquely provided that a definition condition (e.g. of the form $\Phi(0)=0$) is satisfied. Hence it follows that

$$\overline{\Phi(-t)}=\Phi(t), \quad \overline{\Psi(-t)}=\Psi(t). \tag{7}$$

If $\Phi(t)$ and $\Psi(t)$ are expanded into a polynominal series

$$\Phi(t)=\sum_{n=0}^{\infty}\sum_{k=0}^{n}a_{nk}z^k, \quad \Psi(t)=\sum_{n=0}^{\infty}\sum_{k=0}^{n}b_{nk}z^k,$$

then by virtue of (7), we can consider that

$$a_{nk}=(-1)^k\bar{a}_{nk}, \quad b_{nk}=(-1)^k\bar{b}_{nk}. \tag{8}$$

Hence it follows that the coefficients for even powers of t are real, and those for odd powers are purely imaginary. We shall seek the approximations to $\Phi(t)$ and $\Psi(t)$ in the form of fourth power polynomials. By virtue of what has been said above, it is possible to look for them in the form:

$$\Phi(t)=ia_1t+a_2t^2+ia_3t^3+a_4t^4,$$
$$\Psi(t)=a_5+ia_6t+a_7t^2+ia_8t^3+a_9t^4,$$

where the coefficients $a_1, a_2, \ldots, a_9$ are real. Condition (19) § 92 assumes the form:

$$\|ia_1\tau+a_2\tau^2+ia_3\tau^3+a_4\tau^4-2\bar{\tau}\ln|\tau|(-ia_1+2a_2\bar{\tau}-3ia_3\bar{\tau}^2$$
$$+4a_4\bar{\tau}^3)+a_5-ia_6\bar{\tau}+a_7\bar{\tau}^2-ia_8\bar{\tau}^3+a_9\bar{\tau}^4\|^2=\min. \tag{9}$$

We denote by $\varphi_k(\tau)$, $k=1, 2, \ldots, 9$ the coefficient of a_k under the norm sign in (9), so that, for example,

$$\varphi_1(\tau)=i(\tau+2\bar{\tau}\ln|\tau|), \quad \varphi_2(\tau)=\tau^2-4\bar{\tau}^2\ln|\tau|, \tag{10}$$

and so on. Condition (9) assumes the simpler form

$$\left\|\sum_{k=1}^{9}a_k\varphi_k+q\ln|\tau|\right\|^2=\min. \tag{11}$$

Hence, bearing in mind that the coefficients a_k themselves are real, we arrive at the system

$$\sum_{k=1}^{9} a_k \mathrm{Re}\{(\varphi_k, \varphi_j)\} = -q\mathrm{Re}\{(\ln|\tau|, \varphi_j)\}, \quad j=1, 2, \ldots, 9. \quad (12)$$

We introduce the space $L_2(\mathrm{p}; \Gamma)$ with weight function

$$p = \frac{1}{\sqrt{(1+\cos^2 x)}}.$$

Then the evaluation of the integrals which occur in the coefficients of system (12) can be carried out fairly simply with the aid of the well-known formula

$$\frac{1}{\pi}\int_0^{\pi} e^{iz\cos\theta}\cos m\theta \mathrm{d}\theta = i^m J_m(z),$$

which is valid for integral m. Here J_m is Bessel's function of the first kind of order m. Replacing z by in, we reduce this formula to the form

$$\frac{1}{\pi}\int_0^{\pi} e^{-n\cos\theta}\cos m\theta \mathrm{d}\theta = i^m J_m(in).$$

As an example we evaluate the quantity (φ_1, φ_1) which is the coefficient of a_1 in the first equation of system (12). We have

$$(\varphi_1, \varphi_1) = \int_\Gamma (\tau + 2\bar{\tau}\ln|\tau|)(\bar{\tau} + 2\tau\ln|\tau|)\frac{\mathrm{d}s_\tau}{\sqrt{(1+\cos^2 x)}}$$

$$= \int_\Gamma \{|\tau|^2 + 4|\tau|^2\ln^2|\tau| + 2(\tau^2+\bar{\tau}^2)\ln|\tau|\}\frac{\mathrm{d}s_\tau}{\sqrt{(1+\cos^2 x)}}.$$

We introduce x given by $\tau = e^{i(x+i\sin x)}$ as an integration variable. Then

$$(\varphi_1, \varphi_1) = \int_0^{2\pi} e^{-3\sin x}(1 + 4\sin^2 x - 4\sin x\cos 2x)\mathrm{d}x$$

$$= \int_0^{2\pi} e^{-3\sin x}(3 + 2\sin x - 2\cos 2x - 2\sin 3x)\mathrm{d}x.$$

We now make the substitution $x=\frac{\pi}{2}+y$; by virtue of the periodicity of the integrand we can take the limits of integration as $-\pi$ to π:

$$(\varphi_1, \varphi_1)=\int_{-\pi}^{\pi} e^{-3\cos y}(3+2\cos y+2\cos 2y+2\cos 3y)\,dy$$
$$=2\int_{0}^{\pi} e^{-3\cos y}(3+2\cos y+2\cos 2y+2\cos 3y)\,dy$$
$$=2\pi[3J_0(3i)+2iJ_1(3i)-2J_2(3i)-2iJ_3(3i)].$$

We find from the tables of Jahnke and Emde ([1]) that $(\varphi_1, \varphi_1)=2\pi(3\times 4\cdot 8808-2\times 3\cdot 9534+2\times 2\cdot 2452-2\times 0\cdot 9598)$ $=18\cdot 6128\pi$.

In precisely the same way

$$(\varphi_1, \varphi_2)=i\int_{\Gamma}(\tau+2\bar{\tau}\ln|\bar{\tau}|)(\bar{\tau}^2-4\tau^2\ln|\tau|)\frac{ds_\tau}{\sqrt{(1+\cos^2 x)}}$$
$$=\int_{0}^{2\pi} e^{-4\sin x}\{ie^{-ix}-(2ie^{-3ix}-4ie^{3ix})\sin x-8ie^{ix}\sin^2 x\}dx.$$

We now separate the real part:

$$\mathrm{Re}(\varphi_1, \varphi_2)=\int_{0}^{2\pi} e^{-4\sin x}\{\sin x-6\sin 3x\sin x+8\sin^3 x\}dx$$
$$=\int_{0}^{2\pi} e^{-4\sin x}(7\sin x-3\cos 2x+3\cos 4x-2\sin 3x\}dx,$$

which after the substitution $x=\frac{\pi}{2}+y$ gives

$$\mathrm{Re}(\varphi_1, \varphi_2)=2\int_{0}^{\pi} 2e^{-4\cos y}(7\cos y+3\cos 2y+2\cos 3y$$
$$+3\cos 4y)\,dy=2\pi(-7\times 9\cdot 7595+3\times 6\cdot 4222-2\times 3\cdot 3373$$
$$+3\times 1\cdot 4163)=-102\cdot 9512\pi.$$

The remaining coefficients are determined analogously.

Values of $\frac{1}{\pi}(\varphi_k, \varphi_j)$ for equation (12), page 540. The table is arranged symmetrically about the bold line

	1	2	3	4	5	6	7	8	9	Independent term
1	18·613	−102·95	−339·25	897·26	−2·7558	−11·323	28·774	59 119	−110·91	2·01609
2	−102·95	670·49	2492·9	−7343·8	15·162	64·222	−184·47	−440·15	939·67	−14·4819
3	−339 25	2492·9	10485	−34528	40 356	197·78	−668·19	−1847·3	4490·7	−42·9789
4	897·26	−7343·8	−34528	126150	−89·697	−491·45	1904·5	6001·1	−16430	101·579
5	−2·7558	15·162	40·356	−89·697	2·5332	3·1812	−4·4904	−6·6746	10·216	−1·13049
6	−11·323	64·222	197·78	−491·45	3·1812	9·7616	−19·519	−35·011	60·301	−2·96859
7	28·774	−184·47	−668·19	1904·5	−4·4904	−19·519	54·480	122·68	−248·02	4·91329
8	59·119	−440·15	−1847·3	6001·1	−6·6746	−35·011	122·68	337·19	−799·75	7·83859
9	−110·91	939·67	4490·7	−16430	10·216	60·301	−248·02	−799·75	2187·2	12·4899

For the evaluation of Bessel's function of an imaginary argument we have used the tables of Jahnke and Emde [1] and also "Tables of values of Bessel functions of imaginary argument", USSR Acad. Sci. publishers, 1950. In the cases where the necessary values of the argument were lacking in the tables we have used the well known recurrence relation

$$J_{n+1}(x)=J_{n-1}(x)-\frac{2n}{x}J_n(x).$$

On p. 542 we give a table of the coefficients and independent terms of system (12).

Solving system (12) we obtain

$$\begin{aligned}
a_1&=-0{\cdot}6053;\\
a_2&=-0{\cdot}1226;\\
a_3&=\ \ 0{\cdot}007774;\\
a_4&=-0{\cdot}002533;\\
a_5&=\ \ 0{\cdot}01649;\\
a_6&=-0{\cdot}3763;\\
a_7&=\ \ 0{\cdot}2241;\\
a_8&=-0{\cdot}1723;\\
a_9&=-0{\cdot}04594.
\end{aligned}$$

The value of (11) corresponding to the approximation constructed equals $0{\cdot}0159\pi q^2$. We note that

$$\|\ln|\tau|\|^2=\int_0^{2\pi}\sin^2 x e^{-\sin x}\mathrm{d}x=1{\cdot}4018\pi.$$

Hence it is clear that the approximation constructed satisfies the boundary conditions in the mean with an error of

$$\sqrt{\left(\frac{0{\cdot}0159}{1{\cdot}4018}\right)}\approx 10\%.$$

CHAPTER XI

FINITE DIFFERENCE METHODS

§ 94. The method of networks

For the construction of approximate solutions of differential equations wide use is made of so called *finite difference methods.*

The idea underlying all finite difference methods is extremely simple. As is well known a derivative of any order can be obtained as the limit of the ratio of the corresponding "finite difference" of the function to the products of suitable powers of the increments of the independent variables. Thus, for example

$$\frac{\partial u}{\partial x}=\lim_{h\to 0}\frac{u(x+h,\,y,\,z,\,\ldots)-u(x,\,y,\,z,\,\ldots)}{h},$$

$$\frac{\partial^2 u}{\partial x^2}=\lim_{h\to 0}\frac{u(x+h,y,z,\ldots)-2u(x,y,z,\ldots)+u(x-h,y,z,\ldots)}{h^2},$$

$$\frac{\partial^2 u}{\partial x\partial y}=\lim_{h,\,k\to 0}\frac{u(x+h,\,y+k,\,z,\,\ldots)}{hk}$$

$$-\frac{u(x+h,\,y,\,z,\,\ldots)+u(x,\,y+k,\,z,\ldots)-u(x,\,y,\,z,\,\ldots)}{hk}$$

and so on. Hence it follows that for small increments of the independent variables, the derivatives are close to the corresponding ratios. By replacing the derivatives with respect to all or some of the variables by difference ratios we strongly simplify the structure of the given differential equation.

Of the finite difference methods, the method of networks has been most used. We shall elucidate it by one of the simplest

examples—that of the Dirichlet problem for the Laplace equation in a plane. Suppose we are required to find a function which is harmonic in Ω and assumes specified values on the boundary S. We construct a square network on the (x, y) plane by drawing straight lines parallel to the x and y axes separated by a distance h (Fig. 20). We construct a boundary S_h from the sides of the mesh squares which fits the given boundary S as well as possible. The boundary S_h bounds some domain Ω_h which is close to Ω.

We shall seek the values of the unknown harmonic function, not at all points in the domain Ω, but only at the nodes (or mesh vertices) of Ω_h. In precisely the same way, we shall consider the function to be specified not on S but at the nodes of S_h; for this we transfer the given values of the function from S onto S_h by assuming, for example, that at any node P on the boundary S_h the function assumes the same value as at the point P' on the boundary S closest to P. At the points inside Ω_h we replace the partial derivatives by finite differences. Thus if $P(x, y) \in \Omega_h$, then we put

$$\frac{\partial^2 u}{\partial x^2} \approx \frac{(x+h, y)-2u(x, y)+(x-h, y)}{h^2},$$

$$\frac{\partial^2 u}{\partial y^2} \approx \frac{u(x, y+h)-2u(x, y)+u(x, y-h)}{h^2}.$$

The Laplace equation at the point P is replaced by the following algebraic equation:

$$u(x+h, y)+u(x, y+h)+u(x-h, y) + u(x, y-h)-4u(x, y)=0. \tag{1}$$

There are as many such equations as there are points in Ω_h. Equation (1) is homogeneous if all the mesh points neighbouring on P enter into Ω_h and non-homogeneous if some of the neighbouring points of P belong to S_h. Thus, by using the method of networks, the Laplace equation is replaced by a system of an extremely large number of linear algebraic equations.

Instead of a square mesh we can use a rectangular one by making the mesh increments different along the different axes. It is also possible to use polygonal or even curvilinear meshes. It is obvious that the method of networks can be applied to the solution of differential equations of any order with any number of independent variables and unknown functions.

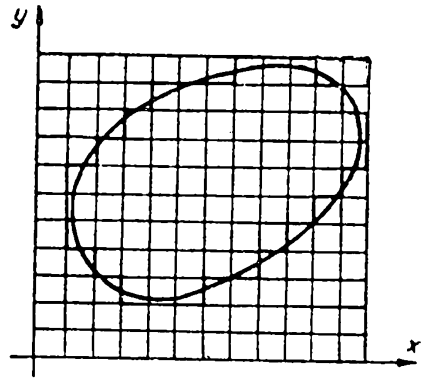

FIG. 20

The actual form of the equation is not important — the method of networks may be successfully applied to equations of elliptic, hyperbolic and parabolic types. The method of networks has given rise to an extensive literature in which many questions both theoretical and practical concerning the method of networks are discussed. Therefore, we shall not touch upon them in this book but merely restrict ourselves to a brief outline of the most important work and to the necessary literature references.

It is difficult to say when the method of networks was first applied. In any case Euler's well-known method for the approximate integration of ordinary differential equations under given initial conditions is essentially a modification of the method of networks. However, a broader development of the method came about in the 20th century. Both the simplicity and the universal nature of the method contributed to its

wide distribution: the form of the domain boundary and the boundary conditions scarcely influence the computational complexity, nor are the properties of the coefficients of the equation of great importance. In engineering practice the method of networks is widely used for calculations concerning elastic systems. We point out in connexion with this the books by G. Markus [1] and P. M. Varvak [1] in which a number of interesting applications of the method of networks are given for problems of elasticity theory. An extensive bibliography is given in the monograph by P. M. Varvak. We should also draw attention to the book by F. Bleikh and Ye. Melan [1]. Extensive material on the method of networks is given in the book by D. Yu. Panov [1]. Some information about the method can also be found in a paper by D. Yu. Panov [2]. The method of networks for elliptic equations is discussed very thoroughly in the book by L. V. Kantorovich and V. I. Krylov [1].

In recent decades the method of networks has been widely used for the solution of a number of complex theoretical problems. The first contribution to this trend was given by L. A. Lyusternik in 1926 in an outstanding paper in which, using the method of networks he gave the solution of the Dirichlet problem in a plane under extremely general assumptions about the nature of the boundary. Incidentally L.A. Lyusternik established the uniform convergence of the approximations obtained by means of the method of networks to the exact solution. (Lyusternik's paper has been reprinted in a new edition and is translated from the Russian in the journal *Uspekhi matem. nauk* cf. L.A. Lyusternik [2].) There is a further paper by the same author [3] in which a method is given for improving the convergence of the method of networks. The case of a general elliptic equation in m variables has been examined by the method of networks in a work by I. G. Petrovsky [1]. In the work of R. Courant, K. Friedrichs and G. Levi [1] the method of networks is applied both to elliptic and to hyperbolic linear equations. The method used here is also readily

extendable to the case of many dimensions. We might also draw attention to the work of Sh. Ye. Mikeladze [1] in which the author gives various estimates of the error associated with the method of networks. In the book by O. A. Ladyzhenskaya [1], and in a number of later works by the same author, the method of networks has been applied to the solution of boundary value problems for hyperbolic and parabolic equations of very general form.

In some of its sections the latest edition of the book by L. V. Kantorovich, V. I. Krylov and K. Ye. Chernin [1] is also concerned with the method of networks.

It was stated above that the method of networks leads (at least in the case of a differential equation of elliptic type) to algebraic systems with a very large number (up to many hundreds) of equations in the same number of unknowns. Obviously the method will be useful in practice if a sufficiently simple method for solving such a system is given. For the system of equations obtained by the method of networks from the Laplace equation, we can construct a simple convergent iterative process. The reader will find a discussion of questions related to this, for example, in the paper by D. Yu. Panov [3]. In the paper by Sh. Ye. Mikeladze [1] mentioned above, a convergent iterative process is constructed for an equation of more general form. Finally, an extremely interesting method for the case of a Laplace or Poisson equation has been indicated in a paper by S. A. Gershgorin [1] which uses a special type of electric network which automatically gives the solution of the systems mentioned above. S. A. Gershgorin's method can be used both for the plane and for the three-dimensional problems. We also mention the paper by M. R. Shura-Bura in which a theoretical estimate using probability theory is given for the error associated with the method of electrical networks.

§ 95. Foundations of the method of straight lines

The method of straight lines which we shall explain in §§ 95–98 bears the same relation to the method of networks as L. V. Kantorovich's method (§ 15) bears to the Ritz method. We shall explain the fundamental ideas behind this method by the example of a second-order equation in two independent variables.

Suppose that we must integrate the equation

$$A\frac{\partial^2 u}{\partial x^2}+2B\frac{\partial^2 u}{\partial x\partial y}+C\frac{\partial^2 u}{\partial y^2}+D\frac{\partial u}{\partial x}+E\frac{\partial u}{\partial y}+Fu=G(x, y) \quad (1)$$

in some domain Ω. To be explicit we suppose that equation (1) is elliptic and that the domain Ω is finite. Some boundary condition must be specified on the boundary S of this domain. We shall assume, although it is not necessary, that the boundary S is such that every straight line parallel to the x-axis meets it in not more than two points.

We draw straight lines parallel to the x-axis and we suppose the distance between two neighbouring lines to be constant and equal to h. Let the straight lines

$$y=y_0+kh=y_k; \quad k=0,\ 1,\ 2,\ \ldots,\ n-1,\ n \quad (2)$$

so constructed intersect the domain Ω. We put $y=y_k$ in equation (1) and replace the derivatives with respect to y by the corresponding finite differences. For example we can put

$$\left.\frac{\partial u}{\partial y}\right|_{y=y_k} \doteq \frac{1}{h}[u(x, y_{k+1})-u(x, y_k)],$$

or, introducing the notation $u(x, y_k)=u_k(x)$,

$$\left.\frac{\partial u}{\partial y}\right|_{y=y_k} = \frac{1}{h}[u_{k+1}(x)-u_k(x)]. \quad (3)$$

Similarly we can put, for example,

$$\left.\frac{\partial^2 u}{\partial x\partial y}\right|_{y=y_k} = \frac{1}{h}[u'_{k+1}(x)-u'_k(x)],$$

$$\left.\frac{\partial^2 u}{\partial y^2}\right|_{y=y_k} = \frac{1}{h^2}[u_{k+1}(x)-2u_k(x)+u_{k-1}(x)].$$

Substituting this in equation (1) with y replaced by y_k we obtain a system of n ordinary linear differential equations in the $n+2$ unknowns $u_0(x)$, $u_1(x)$, ..., $u_n(x)$, $u_{n+1}(x)$.

If the domain Ω has the form depicted in Fig. 21 (p. 554) then the two missing equations can be obtained from the boundary conditions on the straight line segments AB and CD. We shall not pursue this question further in the general case.

The boundary conditions for the functions $u_k(x)$ must be obtained from the boundary conditions for the functions $u(x, y)$.

Analogously it is possible to reduce a higher order linear equation, and also an equation of non-elliptical type, to a system of ordinary differential equations. The method of straight lines can also be readily extended to the case of many dimensions.

The method of straight lines may be considered as a limiting case of the method of networks if we start from a rectangular (not a square) mesh and assume that in the x-direction the mesh step tends to zero. The question of the convergence of the method of straight lines for elliptic equations has been insufficiently investigated. (There is an error in the proof given by M. G. Slobodyansky [2] for the case of Laplace's equation.) The question of replacing the derivatives by difference ratios so as to make the error as small as possible and the question of the most convenient method in practice (for a particular case) for solving the differential equations to which the method of straight lines reduces have been more thoroughly investigated. These questions are discussed in the works of M. G. Slobodyansky and V. N. Faddeyeva [3]. In the following sections we shall formulate the fundamental results obtained by these authors.

We note that the method of straight lines is useful if the coefficients of the given equation are constant or depend only upon y, since in these cases we arrive at a system of linear equations with constant coefficients. If the coefficients of the

given equation depend upon x then the method of straight lines leads to a system of linear equations with variable coefficients the solution of which is, in general, a matter of extreme complexity. In this case, it is more useful to use the methods of networks.

§ 96. Differential equations of the method of straight lines for the equations of Poisson and Laplace

In §§ 96 – 97 we reproduce, almost without change, various sections from the paper by V. N. Faddeyeva [3].

We consider a function $u(y)$ which has a sufficient number of derivatives within some interval.

We shall express the second difference $\Delta^2 u = u(y+2h) - 2u(y+h) + u(y)$ in terms of the second derivatives of u at the points $y+2h$, $y+h$, y to within $0(h^6)$.

By Taylor's theorem we have:

$$u(y+2h) = u(y+h) + h\frac{du(y+h)}{dy} + \frac{h^2}{2}\frac{d^2u(y+h)}{dy^2} + \frac{h^3}{6}\frac{d^3u(y+h)}{dy^3} + \frac{h^4}{24}\frac{d^4u(y+h)}{dy^4} + \dots$$

$$u(y) = u(y+h) - h\frac{du(y+h)}{dy} + \frac{h^2}{2}\frac{d^2u(y+h)}{dy^2} - \frac{h^3}{6}\frac{d^3u(y+h)}{dy^3} + \frac{h^4}{24}\frac{d^4u(y+h)}{dy^4} + \dots .$$

Hence

$$u(y+2h) - 2u(y+h) + u(y) = h^2\frac{d^2u(y+h)}{dy^2} + \frac{h^4}{12}\frac{d^4u(y+h)}{dy^4} + O(h^6). \tag{1}$$

But

$$\frac{d^2u(y+2h)}{dy^2} = \frac{d^2u(y+h)}{dy^2} + h\frac{d^3u(y+h)}{dy^3} + \frac{h^2}{2}\frac{d^4u(y+h)}{dy^4} + \dots$$

$$\frac{d^2u(y)}{dy^2} = \frac{d^2u(y+h)}{dy^2} - h\frac{d^3u(y+h)}{dy^3} + \frac{h^2}{2}\frac{d^4u(y+h)}{dy^4} - \dots$$

and, consequently,

$$h^2 \frac{\mathrm{d}^4 u(y+h)}{\mathrm{d}y^4} = \frac{\mathrm{d}^2 u(y+2h)}{\mathrm{d}y^2} - 2\frac{\mathrm{d}^2 u(y+h)}{\mathrm{d}y^2} + \frac{\mathrm{d}^2 u(y)}{\mathrm{d}y^2} + O(h^4).$$

Substituting the value obtained for the fourth derivative in formula (1), we obtain

$$\Delta^2 u = u(y+2h) - 2u(y+h) + u(y)$$

$$= \frac{5}{6} h^2 \frac{\mathrm{d}^2 u(y+h)}{\mathrm{d}y^2} + \frac{h^2}{12} \frac{\mathrm{d}^2 u(y+2h)}{\mathrm{d}y^2} + \frac{h^2}{12} \frac{\mathrm{d}^2 u(y)}{\mathrm{d}y^2} + O(h^6). \quad (2)$$

This formula will be the starting point for the remainder of the argument.

Suppose we are required to find the function $u(x, y)$ which satisfies the equation

$$\frac{\partial^2 u(x, y)}{\partial x^2} + \frac{\partial^2 u(x, y)}{\partial y^2} = 0 \quad (3)$$

in the domain D under specified boundary conditions on the closed contour bounding it. We assume that the contour is sufficiently smooth and that every straight line parallel to the x-axis intersects it at two points.

We draw straight lines parallel to the x-axis with interval h:

$$y = y_1, \; y = y_2, \; \ldots, \; y = y_n.$$

We use the notation

$$u_k(x) = u(x, \; y_k).$$

Every triplet of functions $u_{k-1}(x)$, $u_k(x)$, $u_{k+1}(x)$ will obviously satisfy relation (2) which we rewrite in the form:

$$u_{k+1}(x) - 2u_k(x) + u_{k-1}(x) = \frac{5}{6} h^2 \left.\frac{\partial u(x, y)}{\partial y^2}\right|_{y=y_k}$$

$$+ \frac{h^2}{12} \left.\frac{\partial^2 u(x, y)}{\partial y^2}\right|_{y=y_{k-1}} + \frac{h^2}{12} \left.\frac{\partial^2 u(x, y)}{\partial y^2}\right|_{y=y_{k+1}} + O(h^6). \quad (2_1)$$

But by virtue of equation (3) we have:

$$\left.\frac{\partial^2 u(x, y)}{\partial y^2}\right|_{y=y_k} = -u_k''(x). \quad (3_1)$$

Substituting the values of the derivatives from formula (3_1) into (2_1) and retaining terms up to order h^6, we obtain the system:

$$\frac{5}{6}u_k''(x)+\frac{1}{12}[u_{k+1}''(x)+u_{k-1}''(x)]$$

$$+\frac{1}{h^2}[u_{k+1}(x)-2u_k(x)+u_{k-1}(x)]=0, \quad k=1, 2, \ldots, n. \quad (4)$$

This system was given by M.G. Slobodyansky [1, 2].

System (4) is a system of ordinary differential equations which, in combination with the boundary conditions, defines the n unknown functions $u_k(x)$. It does not depend upon the form of the boundary of the domain for which the solution is sought. The "boundary" functions $u_0(x)$ and $u_{n+1}(x)$ which appear in it are determined quite naturally from the boundary conditions if the domain boundary contains segments parallel to the x-axis (Fig. 21).

In the case of a curved boundary the question of determining the functions $u_0(x)$ and $u_{n+1}(x)$ requires special investigation. Similarly for the Poisson equation:

$$\nabla^2 u=f(x, y) \quad (5)$$

with the boundary condition $u=0$ we have, on the basis of equation (5),

$$\left.\frac{\partial^2 u(x, y)}{\partial y^2}\right|_{y=y_k}=-u_k''(x)+f_k(x), \quad (6)$$

where $f_k(x)=f(x, y_k)$. Therefore the system for determining the functions $u_k(x)$ has the form:

$$\frac{5}{6}u_k''(x)+\frac{1}{12}[u_{k+1}''(x)+u_{k-1}''(x)]+\frac{1}{h^2}[u_{k+1}(x)-2u_k(x)+u_{k-1}(x)]$$

$$-f_k(x)-\frac{f_{k+1}(x)-2f_k(x)+f_{k-1}(x)}{12}=0, \qquad k=1, 2, \ldots, n,$$

or, collecting like terms

$$\frac{5}{6}u_k''(x)+\frac{1}{12}[u_{k+1}''(x)+u_{k-1}''(x)]+\frac{1}{h^2}[u_{k+1}(x)-2u_k(x)+u_{k-1}(x)]$$

$$-\frac{5}{6}f_k(x)-\frac{f_{k+1}(x)+f_{k-1}(x)}{12}=0, \quad k=1, 2, \ldots, n. \quad (7)$$

By confining ourselves to the determination of the functions $u_k(x)$ i.e. to the determination of the function $u(x, y)$ only on a certain number of straight lines parallel to the x-axis, we obtain for the approximate values of these functions a system of n linear differential equations.

The solution of this system, in general, gives a considerably more accurate result than the solution of the system obtained by the method of finite differences.

§ 97. The case of a trapezoidal domain

In this section we shall consider the more detailed application of the method of straight lines to the integration of Poisson's equation in the case where the domain has the form depicted in Fig 21, i.e. where the domain boundary consists of the segments AB and CD of straight lines parallel to the x-axis and of the arcs AC and BD each of which intersects any straight line parallel to the x-axis in not more than one point.

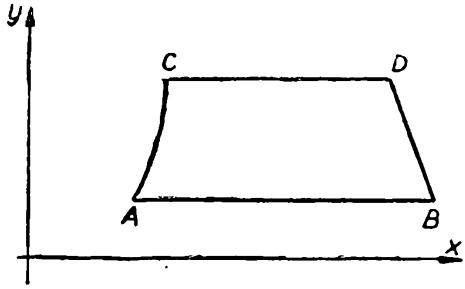

FIG. 21

We consider Poisson's equation

$$\frac{\partial^2 u(x, y)}{\partial x^2}+\frac{\partial^2 u(x, y)}{\partial y^2}=f(x, y)$$

with the boundary condition $u=0$.

As we have already remarked, in this case $u_0(x)=u_{n+1}(x)=u_0''(x)=u_{n+1}''(x)=0$ and therefore we obtain a completely determinate system for finding the $u_k(x)$:

$$\frac{5}{6}u_k''(x)+\frac{1}{12}[u_{k+1}''(x)+u_{k-1}''(x)]$$

$$+\frac{1}{h^2}[u_{k+1}(x)-2u_k(x)+u_{k-1}(x)]-F_k(x)=0, \tag{1}$$

$$k=1, 2, \ldots, n,$$

where

$$F_k(x)=\frac{5}{6}f_k(x)+\frac{f_{k+1}(x)+f_{k-1}(x)}{12}. \tag{2}$$

System (1) is a system of linear differential equations with constant coefficients. We shall transform it into a system each equation of which contains only one unknown function. (The knowledge of linear algebra necessary for the following can be found, for example in the book by V. N. Faddeyeva [4].)

To this end we consider the vectors:

$$\left.\begin{aligned} U&=[u_1(x), u_2(x), \ldots, u_n(x)],\\ F&=[F_1(x), F_2(x), \ldots, F_n(x)].\end{aligned}\right\} \tag{3}$$

We shall denote by A and M the matrices:

$$A=\begin{Vmatrix} \frac{5}{6} & \frac{1}{12} & 0 & 0 & \ldots & 0 & 0\\ \frac{1}{12} & \frac{5}{6} & \frac{1}{12} & 0 & \ldots & 0 & 0\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ 0 & 0 & 0 & 0 & \ldots & \frac{1}{12} & \frac{5}{6} \end{Vmatrix}, M=\begin{Vmatrix} -2 & 1 & 0 & \ldots & 0 & 0\\ 1 & -2 & 1 & \ldots & 0 & 0\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ 0 & 0 & 0 & \ldots & 1 & -2 \end{Vmatrix}.$$

Then system (1) can be rewritten in the form:

$$AU''+\frac{M}{h^2}U-F=0. \tag{4}$$

In order to transform system (1) into a decomposable system, we must reduce the matrices A and M to diagonal form, with the aid of one and the same orthogonal transformation. This is possible. In fact

$$A=E+\frac{M}{12}, \tag{5}$$

where E is the identity matrix. Hence it is clear that the orthogonal transformation to diagonal form of one of the matrices brings about the same transformation of the other matrix.

Thus, let B be the orthogonal transformation which reduces M to diagonal form, i.e. let

$$B^{-1}MB=(\lambda_1, \lambda_2, \ldots, \lambda_n), \text{ where } (\lambda_1, \lambda_2, \ldots, \lambda_n)=$$

$$=\begin{Vmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & \lambda_n \end{Vmatrix}.$$

Then obviously,

$$B^{-1}AB=\left(1+\frac{\lambda_1}{12}, 1+\frac{\lambda_2}{12}, \ldots, 1+\frac{\lambda_n}{12}\right).$$

We now find the orthogonal transformation $B=\|b_{ks}\|$ and the numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$.

We have

$$MB=B(\lambda_1, \lambda_2, \ldots, \lambda_n).$$

Hence by fixing s, we obtain the following system defining the numbers

$$b_{k-1,s}-2b_{k,s}+b_{k+1,s}=b_{ks}\lambda_s; \quad b_0=b_{n+1}=0$$

or

$$b_{k-1,s}+(-2-\lambda_s)b_{k,s}+b_{k+1,s}=0, \quad k=1, 2, \ldots, n. \tag{6}$$

In order to determine all the elements of the matrix B we thus obtain n systems of form (6).

In order that these systems have non-zero solutions, it is necessary that their determinant vanish, i.e. it is necessary that the numbers λ_s be the roots of the characteristic equation:

$$D_n=\begin{vmatrix} -2-\lambda & 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & -2-\lambda & 1 & 0 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \dots & 1 & -2-\lambda \end{vmatrix}=0. \tag{7}$$

We use the notation

$$-2-\lambda=2\cos\theta. \tag{8}$$

Then the determinant D_n can be written in the form

$$D_n=\begin{vmatrix} 2\cos\theta & 1 & 0 & \dots & 0 & 0 \\ 1 & 2\cos\theta & 1 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 1 & 2\cos\theta \end{vmatrix} \tag{9}$$

and can readily be evaluated. Thus

$$D_n=\frac{\sin(n+1)\theta}{\sin\theta}. \tag{10}$$

In order to prove equation (10), we expand D_n about the elements of the first column. This gives the difference equation

$$D_n=2\cos 2\theta D_{n-1}-D_{n-2}.$$

By direct calculation we find

$$D_1=\frac{\sin 2\theta}{\sin\theta}, \quad D_2=\frac{\sin 3\theta}{\sin\theta};$$

and formula (10) is now established by proceeding from n to $n+1$.

The requirement that $D_n=0$ gives n values for θ:

$$\theta_s=\frac{\pi s}{n+1} \quad (s=1,\ 2,\ \dots,\ n) \tag{11}$$

and thus

$$\lambda_s=-2\left(1+\cos\frac{\pi s}{n+1}\right) \quad (s=1,\ 2,\ \dots,\ n). \tag{12}$$

The numbers b_{ks} are the components of the eigenvector of the matrix M for the characteristic values λ_s. They are determined from system (6) to within a constant factor which is then determined from the normalization condition.

We can take the numbers

$$b'_{ks}=\Delta^s_k, \tag{13}$$

as a solution of system (6), where Δ^s_k are the cofactors of the first row of the kth column of the determinant D_n with θ replaced by θ_s.

In fact, by substituting b'_{ks} in the equations of the system, we obtain zeros on the left-hand side of all equations, apart from the kth. These are the sum of the products of the cofactors of one line with the elements of another. The left-hand side of the kth equation transforms into D^s_n, i.e. into the determinant D_n with the substitution of λ_s in place of λ, and is thus also zero.

The cofactors Δ^s_k are readily evaluated.

Thus, let Δ_k be the cofactor of the first row of the kth column of the determinant D_n. Then

$$\Delta=\begin{array}{|cccccc:cccccc|}
\multicolumn{6}{c:}{\overbrace{}^{k-1}} & \multicolumn{6}{c|}{\overbrace{}^{n-k}} \\
\hline
1 & 2\cos\theta & 1 & 0\ldots 0 & 0 & & 0 & 0 & 0\ldots 0 & 0 \\
0 & 1 & 2\cos\theta & 1\ldots 0 & 0 & & 0 & 0 & 0\ldots 0 & 0 \\
\multicolumn{6}{c:}{\cdots\cdots\cdots\cdots\cdots} & \multicolumn{4}{c}{\cdots\cdots\cdots\cdots} \\
0 & 0 & 0 & 0\ldots 1 & 2\cos\theta & & 0 & 0 & 0\ldots 0 & 0 \\
0 & 0 & 0 & 0\ldots 0 & 1 & & 1 & 0 & 0\ldots 0 & 0 \\
\hline
0 & 0 & 0 & 0\ldots 0 & 0 & & 2\cos\theta & 1 & 0\ldots 0 & 0 \\
0 & 0 & 0 & 0\ldots 0 & 0 & & 1 & 2\cos\theta & 1\ldots 0 & 0 \\
\multicolumn{6}{c:}{\cdots\cdots\cdots\cdots\cdots} & \multicolumn{4}{c}{\cdots\cdots\cdots\cdots} \\
0 & 0 & 0 & 0\ldots 0 & 0 & & 0 & 0 & 0\ldots 1 & 2\cos\theta
\end{array}.$$

Hence, by expanding Δ_k in terms of the elements of the first column $k-1$ times, we obtain

$$\Delta^k=(-1)^{k-1}D_{n-k}=(-1)^{k-1}\frac{\sin(n+1-k)\theta}{\sin\theta}.$$

This gives

$$\Delta_k^s = (-1)^{k-1} \frac{\sin(n+1-k)\theta_s}{\sin\theta_s}. \tag{14}$$

Finally we normalize the elements of the matrix B by columns, i.e. we define c_s such that

$$c_s^2 \sum_{k=1}^{n} b_{ks}'^2 = c_s^2 \sum_{k=1}^{n} \frac{\sin^2(n+1-k)\theta_s}{\sin^2\theta_s} = 1.$$

It is readily calculated that

$$\sum_{k=1}^{n} \sin^2(n+1-k)\theta_s = \frac{n+1}{2}.$$

This gives

$$c_s = \sqrt{\left(\frac{2}{n+1}\right)} \sin\theta_s. \tag{15}$$

Thus the elements of the transformation matrix B are:

$$b_{ks} = b_{ks}' c_s = (-1)^{k-1} \sqrt{\left(\frac{2}{n+1}\right)} \sin(n+1-k)\frac{\pi s}{n+1}$$

$$= (-1)^{k+s} \sqrt{\left(\frac{2}{n+1}\right)} \sin\frac{\pi s k}{n+1}. \tag{16}$$

Hence it is clear that the matrix B is symmetric, and, since it is also orthogonal, that

$$B = B^* = B^{-1} = \|b_{ks}\| = \left\| (-1)^{s+k} \sqrt{\left(\frac{2}{n+1}\right)} \sin\frac{\pi s k}{n+1} \right\|. \tag{17}$$

Thus,

$$M = B(\lambda_1, \lambda_2, \ldots, \lambda_n)B,$$

$$A = B\left(1+\frac{\lambda_1}{12}, 1+\frac{\lambda_2}{12}, \ldots, 1+\frac{\lambda_n}{12}\right)B,$$

and therefore, system (4) will have the form:

$$\left(1+\frac{\lambda_1}{12}, \ldots, 1+\frac{\lambda_n}{12}\right) BU'' + \frac{1}{h_2}(\lambda_1, \ldots, \lambda_n)BU - BF = 0.$$

We shall use the notation

$$BU = V = (v_1, v_2, \ldots, v_n),$$
$$BF = G = (g_1, g_2, \ldots, g_n). \tag{18}$$

Then

$$\left(1+\frac{\lambda_1}{12}, \ldots, 1+\frac{\lambda_n}{12}\right) V'' + \frac{1}{h^2}(\lambda_1, \ldots, \lambda_n)V - G = 0.$$

The last equation is equivalent to the n independent equations

$$\left(1+\frac{\lambda_k}{12}\right) v_k''(x) + \frac{\lambda_k}{h^2} v_k(x) - g_k(x) = 0, \quad k = 1, 2, \ldots, n. \tag{19}$$

We now introduce the notation

$$\alpha_k^2 = -\frac{\lambda_k}{h^2\left(1+\frac{\lambda_k}{12}\right)} = \frac{12(1+\cos\theta_k)}{h^2(5-\cos\theta_k)} \tag{20}$$

and

$$\varphi_k(x) = \frac{g_k(x)}{1+\frac{\lambda_k}{12}} = \frac{6g_k(x)}{5-\cos\theta_k}. \tag{21}$$

Then system (19) can be rewritten in the form

$$v_k''(x) - \alpha_k^2 v_k(x) - \varphi_k(x) = 0. \tag{19_1}$$

We shall assume for the sake of simplicity that both the domain boundary and the boundary conditions are symmetric about the x-axis. The solution of the system will be an even function of x and, obviously, $v_k = C_k \cosh \alpha_k x + \bar{v}_k(x)$, where $\bar{v}_k(x)$ is an even particular integral of system (19_1).

It is expedient to find the particular integral of system (19_1) directly (for example, by the method of undetermined multipliers) in the case where the functions $\varphi_k(x)$ have a sufficiently simple form.

It is possible to represent the particular solution of system (19_1) in the form:

$$\bar{v}_k(x) = \frac{1}{\alpha_k}\int_0^x \sinh \alpha_k(x-t)\varphi_k(t)\,dt. \tag{22}$$

Transforming back to the functions $u_k(x)$ we obtain, on the basis of (18),

$$U = B^{-1}V = BV$$

or

$$u_k(x) = \sum_{s=1}^{n} b_{ks} v_s(x). \tag{23}$$

Hence

$$u_k(x) = \sum_{s=1}^{n} b_{ks} C_s \cosh \alpha_s x + \sum_{s=1}^{n}{}' b_{ks} \bar{v}(x)$$

$$= \sum_{s=1}^{n} (-1)^{k+s} C_s \sqrt{\left(\frac{2}{n+1}\right)} \sin \frac{\pi k s}{n+1} \cosh \alpha_s x$$

$$+ \int_0^x \sum_{s=1}^{n} \frac{(-1)^{k+s} \sqrt{(2)} \sin \dfrac{\pi k s}{n+1}}{\sqrt{(n+1)\alpha_s}} \sinh \alpha_s (x-t) \varphi_s(t) \mathrm{d}t. \tag{24}$$

We note here that the nature of solution (24) does not depend upon the form of the domain boundary. The constants $C_1, C_2, \ldots, C_n$ are determined from the algebraic system obtained from (24) by substituting the boundary conditions in it.

For convenience in computation it is expedient to put

$$A_s = C_s \sqrt{\left(\frac{2}{n+1}\right)} \cosh \alpha_s x_m,$$

where x_m is the greatest of the abscissae of the points of intersection of the straight lines $y = y_k$ with the boundary of the domain D.

Then the solution may be written in the form:

$$u_k(x) = \sum_{s=1}^{n} (-1)^{k+s} \sin \frac{\pi k s}{n+1} A_s \frac{\cosh \alpha_s x}{\cosh \alpha_s x_m}$$

$$+ \int_0^x \sum_{s=1}^{n} \frac{(-1)^{k+s} \sin \dfrac{\pi k s}{n+1}}{\sqrt{\left(\dfrac{n+1}{2}\right)\alpha_s}} \sinh \alpha_s (x-t) \varphi_s(t) \mathrm{d}t. \tag{25}$$

We also note that sometimes the particular integral of system (1) can be found immediately. Thus, for example, in the torsion problem

$$\frac{\partial^2 u(x, y)}{\partial x^2}+\frac{\partial^2 u(x, y)}{\partial y^2}=-1 \tag{26}$$

with $u=0$ on the domain boundary system, (1) will have the form:

$$\frac{5}{6}u_k''(x)+\frac{1}{12}[u_{k+1}''(x)+u_{k+1}''(x)]$$

$$+\frac{1}{h^2}[u_{k+1}(x)-2u_k(x)+u_{k-1}(x)]+1=0 \tag{27}$$

and thus the numbers a_k determined from the system

$$a_{k-1}-2a_k+a_{k+1}+h^2=0 \tag{28}$$

will be particular integrals of it.

It is readily seen that

$$a_k=\frac{k(n+1-k)}{2}h^2. \tag{29}$$

In this case the required functions $u_k(x)$ will be evaluated from the formula

$$u_k(x)=\sum_{s=1}^{n}(-1)^{k+s}\sin\frac{\pi ks}{n+1}A_s\frac{\cosh\alpha_s x_k}{\cosh\alpha_s x_m}+\frac{k(n+1-k)}{2}h^2, \tag{30}$$

and our problem will merely consist of finding the constants A_s from the linear algebraic system:

$$\sum_{s=1}^{n}(-1)^{k+s}\sin\frac{\pi ks}{n+1}A_s\frac{\cosh\alpha_s x_k}{\cosh\alpha_s x_m}+\frac{k(n+1-k)}{2}h^2=0, \tag{31}$$

which we obtain by satisfying the boundary conditions at the points on the boundary ordinates y_k.

The numbers α_s, $\delta_k^s=(-1)^{k+s}\sin\dfrac{\pi ks}{n+1}$, $a_k=\dfrac{k(n+1-k)}{2}$

depend only upon the number of straight lines $y=y_k$ that we draw. In the paper by V. N. Faddeyeva tables of these quantities are given for the values $n=3, 5, 7, 11, 15$.

§ 98. Differential equations of the method of straight lines for the biharmonic equation

The equations mentioned in the section title are obtained most readily by starting with the well-known representation of any biharmonic function $u(x, y)$ in terms of two harmonic functions $\varphi(x, y)$ and $\psi(x, y)$

$$u(x, y)=\varphi(x, y)+y\psi(x, y). \tag{1}$$

We obtain the result required by merely writing down the equations of the method of straight lines (4) § 96 for the functions $\varphi(x, y)$ and $\psi(x, y)$. This gives the following system of differential equations with constant coefficients:

$$\begin{aligned}
&\frac{5}{6}\varphi_k''(x)+\frac{1}{12}[\varphi_{k-1}''(x)+\varphi_{k+1}''(x)]\\
&+\frac{1}{h^2}[\varphi_{k-1}(x)-2\varphi_k(x)+\varphi_{k+1}(x)]=0,\\
&\frac{5}{6}\psi_k''(x)+\frac{1}{12}[\psi_{k-1}''(x)+\psi_{k+1}''(x)]\\
&+\frac{1}{h^2}[\psi_{k-1}(x)-2\psi_k(x)+\psi_{k+1}(x)]=0, \quad k=1, 2, \ldots, n.
\end{aligned} \tag{2}$$

System (2) contains $2n$ equations in the $2n+4$ unknowns

$$\varphi_0, \varphi_1, \ldots, \varphi_n, \varphi_{n+1}; \psi_0, \psi_1, \ldots, \psi_n, \psi_{n+1}.$$

We shall suppose that the domain has the form depicted in Fig. 21 and that $H=(n+1)h$ is its height. We further suppose that on the domain boundary the values of the function $u(x, y)$ itself and of its normal derivative are specified. Then on the straight line AB the quantities

$$[\varphi(x, y)+y\psi(x, y)]_{y=0}=\varphi_0(x), \tag{3}$$

$$\frac{\partial}{\partial y}[\varphi(x,y)+y\psi(x,y)]_{y=0}=\left.\frac{\partial\varphi}{\partial y}\right|_{y=0}+\psi_0(x), \tag{4}$$

are specified, and on the straight line CD — the quantities

$$[\varphi(x, y)+y\psi(x, y)]_{y=H}=\varphi_{n+1}(x)+H\psi_{n+1}(x), \tag{5}$$

$$\frac{\partial}{\partial y}[\varphi(x, y)+y\psi(x, y)]_{y=H}=\left.\frac{\partial\varphi}{\partial y}\right|_{y=H}+H\left.\frac{\partial\psi}{\partial y}\right|_{y=H}+\psi_{n+1}(x). \tag{6}$$

We shall find an expression for the quantity $\left.\frac{\partial\varphi}{\partial y}\right|_{y=0}$ and its analogues in terms of the functions $\varphi_k(x)$ and $\psi_k(x)$. By Taylor's theorem

$$\varphi(x, y+l)=\varphi(x, y)+l\frac{\partial\varphi}{\partial y}+\frac{l^2}{2}\frac{\partial^2\varphi}{\partial y^2}+\frac{l^3}{6}\frac{\partial^3\varphi}{\partial y^3}+O(l^4), \tag{7}$$

where $0(\xi)$ denotes a quantity of the same order of magnitude as ξ. We put $y=y_k$, $l=h$ in (7). Then

$$\varphi_{k+1}=\varphi_k(x)+h\left(\frac{\partial\varphi}{\partial y}\right)_{y=y_k}+\frac{h^2}{2}\left(\frac{\partial^2\varphi}{\partial y^2}\right)_{y=y_k}+\frac{h^3}{6}\left(\frac{\partial^3\varphi}{\partial y^3}\right)_{y=y_k}+O(h^4). \tag{8}$$

But $\varphi(x, y)$ is a harmonic function, therefore

$$\left(\frac{\partial^2\varphi}{\partial y^2}\right)_{y=y_k}=-\left(\frac{\partial^2\varphi}{\partial x^2}\right)_{y=y_k}=-\varphi_k''(x),$$

$$\left(\frac{\partial^3\varphi}{\partial y^3}\right)=-\frac{\partial}{\partial y}\left(\frac{\partial^2\varphi}{\partial x^2}\right)=-\frac{\partial^2}{\partial x^2}\left(\frac{\partial\varphi}{\partial y}\right)$$

and, consequently,

$$\left(\frac{\partial^3\varphi}{\partial y^3}\right)_{y=y_k}=-\frac{\partial^2}{\partial x^2}\left(\frac{\partial\varphi}{\partial y}\right)_{y=y_k}.$$

Substituting this in (8) we obtain

$$\varphi_{k+1}(x)=\varphi_k(x)+h\left(\frac{\partial\varphi}{\partial y}\right)_{y=y_k}-\frac{h^2}{2}\varphi_k''(x)-\frac{h^3}{6}\frac{\partial^2}{\partial x^2}\left(\frac{\partial\varphi}{\partial y}\right)_{y=y_k}+O(h^4). \tag{9}$$

We drop the term

$$-\frac{h^3}{6}\frac{\partial^2}{\partial x^2}\left(\frac{\partial\varphi}{\partial y}\right)$$

from (9) and, from the equation so obtained, we find the quantity

$$\left(\frac{\partial\varphi}{\partial y}\right)_{y=y_k} = \frac{\varphi_{k+1}(x)-\varphi_k(x)}{h} + \frac{h}{2}\varphi_k'' + O(h^2). \tag{10}$$

Substituting this in the fourth term on the right in (9) and dropping terms of the order of h^4, we obtain

$$\left(\frac{\partial\varphi}{\partial y}\right)_{y=y_k} = \frac{\varphi_{k+1}(x)-\varphi_k(x)}{h} + \frac{h}{6}\varphi_{k+1}''(x) + \frac{h}{3}\varphi_k''(x) + O(h^3).$$

Hence, for $k=0$, we obtain the approximate formula

$$\left(\frac{\partial\varphi}{\partial y}\right)_{y=0} = \frac{\varphi_1(x)-\varphi_0(x)}{h} + \frac{h}{6}\varphi_1''(x) + \frac{h}{3}\varphi_0''(x). \tag{11}$$

Substituting $y=y_k$, $l=-h$ in (7) we find

$$\left(\frac{\partial\varphi}{\partial y}\right)_{y=y_k} = \frac{\varphi_k(x)-\varphi_{k-1}(x)}{h} - \frac{h}{6}\varphi_{k-1}''(x) - \frac{h}{3}\varphi_k''(x),$$

which for $k=n+1$ gives

$$\left(\frac{\partial\varphi}{\partial y}\right)_{y=H} = \frac{\varphi_{n+1}(x)-\varphi_n(x)}{h} - \frac{h}{6}\varphi_n''(x) - \frac{h}{3}\varphi_{n+1}''(x). \tag{12}$$

Equation (3) defines the quantity $\varphi_0(x)$. If now we substitute the expressions for

$$\left(\frac{\partial\varphi}{\partial y}\right)_{y=0}, \quad \left(\frac{\partial\varphi}{\partial y}\right)_{y=H}, \quad \left(\frac{\partial\psi}{\partial y}\right)_{y=H},$$

obtained from formulae (11) and (12) into (4) and (6), we then obtain three differential equations in the unknowns $\varphi_1(x)$, $\varphi_n(x)$, $\varphi_{n+1}(x)$; $\psi_n(x)$, $\psi_{n+1}(x)$. Combining these equations with equations (2) we obtain a system of $2n+3$ differential equations in the $2n+3$ unknowns $\varphi_1, \ldots, \varphi_{n+1}, \psi_0, \psi_1, \ldots, \psi_{n+1}$. We obtain the boundary conditions for this system by using the boundary conditions of the problem at the points of intersection of the domain boundary with the straight lines $y=y_k$.

Another method of constructing the equations of the method of straight lines is given in a paper by A. Langenbakh [1]; this method gives a system whose solution can be written down immediately.

It is possible to replace the biharmonic equation approximately by the system

$$u_k^{(4)} + \frac{1}{6}\Delta^2 u_k^{(4)} + \frac{2}{h^2}\Delta^2 u_k'' + \frac{1}{6h^2}\Delta^4 u_k'' + \frac{1}{h^4}u_k = 0, \quad k = 1, 2, \ldots, n \tag{13}$$

given by M. G. Slobodyansky [1], [2] in the same way as this was done in § 96 for the Laplace equation. Here

$$\Delta^2 u_k = u_{k+1} - 2u_k + u_{k-1}, \Delta^4 u_k = u_{k+2} - 4u_{k+1} + 6u_k - 4u_{k-1} + u_{k-2}$$

Let M denote the same matrix as in § 97. System (3) can be written in the form

$$\left(E + \frac{1}{6}M\right)U^{(4)} + \frac{2}{h^2}\left(M + \frac{1}{2}M^2\right)U'' + \frac{1}{h^4}M^2U = F, \tag{14}$$

where E is the identity matrix, U is a column vector composed of the functions $u_1(x)$, $u_2(x)$, ..., $u_n(x)$, F is some vector defined in terms of the boundary data and in terms of the yet unknown values of u_1 and u_n. We now put approximately

$$u_1(x) = g_{11}(x) + hg_{21}(x), \quad u_n(x) = g_{12}(x) + hg_{22}(x),$$

in F, where g_{11} and g_{12} denote the given values of the required biharmonic function $u(x, y)$ on the straight line segments AB and CD (Fig. 21) and g_{21} and g_{22} are the values of the normal derivatives of the function $u(x, y)$ on the same segments. Now that the vector F is known, we put (retaining the notation of § 97)

$$BU = V = (v_1, v_2, \ldots, v_n), \quad BF = G = (g_1, g_2, \ldots, g_n);$$

this gives us the n independent equations

$$v_k^{(4)} + \alpha_k v_k'' + \beta_k v_k = \frac{g_k}{1 + \frac{1}{6}\lambda_k}, \quad k = 1, 2, \ldots, n. \tag{15}$$

Here

$$\alpha_k = \frac{2}{h^2}\frac{\lambda_k + \frac{1}{12}\lambda_k^2}{1 + \frac{1}{6}\lambda_k}; \qquad \beta_k = \frac{1}{h^4}\frac{\lambda_k^2}{1 + \frac{1}{6}\lambda_k}.$$

BIBLIOGRAPHY

AKHIYEZER N. I. and I. M. GLAZMAN

1. *Theory of linear operators (Teoriya lineinikh operatorov)*, Gostekhizdat (1950).

BERGMAN S. and M. SCHIFFER

1. *Kernel functions and elliptic differential equations in mathematical physics*, Academic Press, New York (1953).

BIRMAN M. SH.

1. Estimates in the method of steepest descents, *Usp. matem. nauk* **5**, 3 (37), (1950).

2. The calculation of eigenvalues by the method of steepest descents, *Zap. Leningr. gorn. in-ta*, **27**, 1 (1952).

3. Minimal functionals for elliptic differential equations of the second order, *Dokl. Akad. Nauk SSSR*, **93**, 6 (1953).

4. The spectrum of elliptic differential equations with singular boundary conditions, *Dokl. Akad. Nauk SSSR*, **97**, 1 (1954).

5. Trefftz's variational method for the equation $\nabla^4 u = f$, *Dokl. Akad. Nauk SSSR*, **101**, 2 (1955).

6. Variational methods of solving boundary value problems similar to the Trefftz method, *Vest. Len. gos. un-ta, Seriya matem mekh. i astr.* 3, (1956).

BLEIKH F. and YE. MELAN

1. *The finite difference equations of the statics of structures (Uravneniya v konechnykh raznostyakh statiki sooruzhenii)*, United Scientific and Technical Press (ONTI DNTVU NKTP), Khar'kov (1936).

BUBNOV I. G.

1. Report on the works of Prof. Timoshenko which were awarded the Zhuranskii prize. Symposium of the Institute of Communication Engineers *(Sborn. inta inzh. putei soobshch.)*, No. 81, All Union Special Planning Office (SPB) (1913).

CHAPLYGIN S. A.

1. Gas jets (O gasovykh struyakh). In the collection: *Selected works on mechanics and mathematics (izbrannye trudy po mekhanike i matematike)*, Gostekhizdat (1954).

COURANT R.

1. Über ein Konvergenzerzeugen des Prinzips in der Variationsrechnung, *Gött. Nachrichten* (1922).

COURANT R. and D. HILBERT

1. *Methoden der mathematischen Physik I.* English translation (with additions by R. Courant) published as *Methods of Mathematical Physics* by Interscience, New York (1953).

2. *Methoden der mathematischen Physik II,* English translation published as *Methods of Mathematical Physics* by Interscience, New York (1962).

COURANT R., K. FRIEDRICHS and G. LEVI

1. Difference equations of mathematical physics, *Usp. Matem. Nauk,* VIII, 125–60 (1940).

DIAZ J. B.

1. Upper and lower bounds for quadratic functionals, *Proc. Sympos. Spectral theory and differential problems, Stillwater, Oklahoma,* 279–289 (1955).

EIDUS D. M.

1. Mixed problems in the theory of elasticity, *Dokl. Akad. Nauk SSSR,* **76,** 2 (1951).

2. Contact problems in the theory of elasticity, *Matem. sborn.,* **54,** 3 (1954).

FADDEYEVA V. N.

1. On fundamental functions of the operator X^{IV} (O fundamental'nykh funktsiyakh operatora X^{IV}), *Trudy LIIPS* No. 6 (1935).

2. On fundamental functions of the operator X^{IV}, *Trudy matem. in-ta Akad. Nauk SSSR* im. V. A. Steklova, Vol. 28 (1949).

3. The method of lines applied to some boundary problems, *Trudy matem. in-ta Akad. Nauk SSSR* im. V. A. Steklova, Vol. 28 (1949).

4. *Computational methods of linear algebra (Vychislitel'nye metody lineinoi algebry),* Gostekhizdat (1950).

FIKHTENGOL'TS G. M.

1. *The Fundatmentals of Mathematical Analysis (Osnovy matematiches kogo analiza),* Vol. 1, Gostekhizdat (1955). English translation of Vols. 1 and 2 in preparation by Pergamon Press.

FRIEDRICHS K.

1. Randwert- und Eigenwertprobleme aus der Theorie der elastischen Platten, *Math. Ann.,* **98** (1928).

2. Spektraltheorie halbbeschränkter Operatoren und Anwendung auf die Spektralzerlegung von Differentialoperatoren, *Math. Ann.,* **109,** 4–5 (1934).

3. On the boundary value of the theory of elasticity and Korn's inequality, *Annals of Mathematics,* **48,** 2 (1947).

FUCHS B. A. and B. V. SHABAT

1. *Functions of a complex variable and some of their applications, (Funktsii kompleksnogo peremennogo i nekotorye ikh prilozheniya)*

Gostekhizdat (1949). English translation published by Pergamon Press, Oxford (1964).

GALERKIN B. G.

1. Rods and plates. Series occuring in various questions concerning the elastic equilibrium of rods and plates. (Sterzhni i plastiny. Ryady v nekotorykh voprosakh uprogogo ravnovesiya sterzhnei i plastin.) *Engineers Bulletin (Vestnik inzhenerov)* 19, 897–908 (1915).

GEL'FAND I. M.

1. *Lectures on linear algebra (Lektsii po lineinoi algebre)*, Gostekhizdat (1951).

GERSHGORIN S. A.

1. Approximate solutions of Laplace's differential equation by electric networks, *Zhurn. prikl. fiz.* **6**, 3—4 (1929).

GÜNTHER N. M.

1. *La théorie du potentiel et ses applications aux problèmes fondamentaux de la physique mathematique*, Gauthier–Villars, Paris (1934).

JAHNKE E. and F. EMDE

1. *Tables of functions*, Dover, New York (1943).

KANTOROVICH L. V.

1. Functional analysis and applied mathematics, *Usp. matem. nauk*, **111**, 6 (28) (1948).

KANTOROVICH L. V. and V. I. KRYLOV

1. *Approximate methods of higher analysis (Priblizhennye metody vysshego analiza)*, Gostekhizdat (1949). English translation published by P. Noordhoff, Groningen (1958).

KANTOROVICH L. V., V. I. KRYLOV and K. YE. CHERNIN

1. *Tables for the numerical solution of boundary value problems in the theory of harmonic functions (Tablitsy dlya chislennogo resheniya kraevykh zadach teorii garmonicheskikh funktsii)*, Gostekhizdat (1956).

KATO T.

1. On some approximate methods concerning the operators T^*T, *Math. Ann.* **126**, 3, 253–262 (1953)

KELDYSH M. V.

1. The solution of boundary value problems by Galerkin's method, *Izv. Akad. Nauk SSSR, ser. matem.*, **6**, 6 (1942).

2. Some problems of degenerate equations of elliptic type, *Dokl. Akad. Nauk SSSR* **77**, 2 (1951)

KHARRIK I. YU.

1. The approximation to singular functions by functions which are zero on the boundary of a domain, *Matem. sborn.* **37** (79), 2 (1955).

KIRCHHOFF G.

1. Über die Gleichgewicht und die Bewegung einer elastischen Scheibe, *J. reine angew. Mathematik*, **40** (1850)

KLYOT-DASHINSKY M. I.

1. The solution of the problem of the static deformation of an anisotropic homogeneous plate by the method of orthogonal projections, *Uch. zap. Len. gos. un-ta*, No. 146, *ser. fiz. nauk*, 8, 131–145 (1952).

2. A particular method of solving plane potential theory problems, *Sb. nauch. trudov Leningrad. inzh-stroit. in-ta*, 17 (1954).

KRASNOSEL'SKY M. A.

1. Convergence of Galerkin's method for non-linear equations, *Dokl. Akad. Nauk SSSR* **73**, 6 (1950).

2. Some problems of non-linear analysis, *Usp. matem. nauk* IX, 3 (61) (1954).

3. *Topological methods in the theory of nonlinear integral equations (Topologicheskie metody v teorii nelineinykh integral'nykh uravnenii)*, Gostekhizdat (1956). English translation published by Pergamon Press, Oxford (1963).

KRAVCHUK M.

1. Resolution of linear differential equations and integrals by the method of moments, *Vidavn. Ukr. Akad. Nauk Kiev* (1936).

2. Sur la résolution approchée des équations intégrales linéaires, *C. R. t.* **188**, 978 (1929). (vide Krawtchouk).

LADYZHENSKAYA O. A.

1. *The mixed problem for a hyperbolic equation (Smeshannaya zadacha dlya giperbolicheskogo uravneniya)*, Gostekhizdat (1953).

LANGENBAKH A.

1. The approximate solution of the biharmonic equation in a trapezoidal domain, *Vest. Len. gos. un-ta*, 13, *ser. matem, mekh. i astr.* 3 (1956).

LEIBENZON L. S.

1. *Variational methods for solving problems of elasticity theory (Variatsionnye metody resheniya zadach teorii uprugosti)*, Gostekhizdat (1943). See also L. S. Leibenzon, Collected Works Vol. 1, *Izd. Akad. Nauk SSSR* (1951).

2. *Course on the theory of elasticity (Kurs teorii uprugosti)*, Gostekhizdat (1947).

LEVIN V. I. and YU. I. GROSBERG

1. *The differential equations of mathematical physics (Differentsial'nye uravneniya matematicheskoi fiziki)*, Gostekhizdat (1951).

LIN HUNG-SUN

1. On variational methods in the method of the torsion for multiply-connected cross-sections, *Acta Sci-Sinica* 3, 171–186 (1954). (Russian translation from Chinese with an English summary.)

LYUSTERNIK L. A.

1. Über einige Anwendungen der direkten Methoden in Variationsrechnung, *Matem. sborn.* 33, 189–200 (1926).

2. The Dirichlet problem, *Usp. Matem. Nauk*, 8, 115–124 (1940).

3. A remark on the numerical solution of boundary value problems in Laplace's equation and the calculation of eigenvalues by the method of networks, *Trud. Matem. in-ta*, im. V. A. Steklova, **20**, 49–64 (1947).

MAKHOVER YE. V.

1. The bending of plates of variable thickness with sharp edges, *Uch. zap. Len. Gos. Ped. in-ta fiz-matem. f-ta*, **17**, 2, 28–39.

MARKUS G.

1. *The theory of elastic frameworks and its application to calculations connected with plates and pillarless spans (Teoriya uprugoi setki i eye prilozhenie k raschetu plit i bezbalochnykh perekrytii)* DNTVU, Kiev (1936).

MIKELADZE SH. YE.

1. The numerical integration of equations of elliptic and hyperbolic type, *Izv. Akad. Nauk SSSR, ser. matem.* 5, 1 (1941).

MIKHLIN S. G.

1. The plane problem of the theory of elasticity for an inhomogeneous medium, *Trud. Seismolog. in-ta, Akad. Nauk SSSR*, 66 (1935).

2. On a theorem by F. Noether, *Dokl. Akad. Nauk SSSR*, **43**, 4 (1944).

3. The convergence of the method of least squares, *Dokl. Akad. Nauk SSSR* **59**, 7 (1948).

4. The convergence of Galerkin's method, *Dokl. Akad. Nauk SSSR*, **61**, 2 (1948).

5. Singular integral equations, *Usp. Matem. Nauk* **3**, 3 (1948).

6. The method of least squares in problems of mathematical physics, *Uch. zap. Len. gos. un-ta*, 111, *ser. matem. nauk*, **16** (1954).

7. *Integral equations (Integral'nyye uravneniya)*, 2nd edition, Gostekhizdat (1949). English translation published by Pergamon Press, Oxford (1957).

8. Some sufficient conditions for the convergence of Galerkin's method, *Uch. zap. Len. gos. un-ta*, 135, *ser. matem. nauk*, 21, (1950).

9. Some theorems in the theory of operators and their application to the theory of elastic shells, *Dokl. Akad. Nauk SSSR*, **84**, 5 (1952).

10. Estimate of error in calculations replacing an elastic shell by a plane plate, *Prikl. matem. i mekh.*, **16**, 4 (1952).

11. *The problem of the minimum of a quadratic functional (Problema minimuma kvadratichnogo funktsionala)*, Gostekhizdat, (1952).

12. Variational methods in solving boundary value problems *Uch. zap. Len. gos. un-ta*, 144, *ser. matem. nauk*, 3 (1952).

13. Application of variational methods to degenerate elliptic equations, *Dokl. Akad. Nauk SSSR*, **91**, 4 (1953).

14. Integration of Poisson's equation over an infinite domain, *Dokl. Akad. Nauk SSSR*, **91**, 5 (1953).

15. The theory of degenerate elliptic equations, *Dokl. Akad. Nauk SSSR*, **94**, 2 (1954).

16. Degenerate elliptic equations, *Vest. Len. gos. un-ta*, 8 (1954).

17. Elastic envelopes which are nearly plane discs. Symposium: *Questions on the durability of water turbine blades* (Voprosy prochnosti lopasti vodyanoi turviny), *Izd. Len. gos. un-ta*, 5–92 (1954).

18. Foundation of the Ritz method, *Dokl. Akad, Nauk SSSR*, **106**, 3 (1956).

MOREN K.

1. Remarks on the Trefftz and Ritz methods, *Bulletin Polish Acad. Sci.*, Section 3, **3**, No. 11, 569–573 (1955).

MUSKHELISHVILI N. I.

1. *Singular integral equations (Singulyarnyye integral'nnye uravneniya)*, Gostekhizdat (1946). English translation published by P. Noordhoff, Groningen (1953).

2. *Some basic problems of the mathematical theory of elasticity (Nekotoryye osnovnyye Zadachi matematicheskoi teorii uprugosti)*, Izd. Akad. Nauk SSSR (1949). English translation published by P. Noordhoff, Groningen (1953).

NATANSON I. P.

1. *Theory of functions of a real variable (Teoriya funkstii veshchestvennoi peremennoi)*, Gostekhizdat (1950).

NOVOZHILOV V. V.

1. *Foundations of non-linear elasticity theory (Osnovy nelineinoi teorii uprugosti)*, Gostekhizdat (1948).

OLEINIK O. A.

1. Equations of elliptic type degenerating of the boundary of a domain, *Dokl. Akad. Nauk SSSR*, **87**, 6 (1952).

PANOV D. YU.

1. *Handbook for the numerical solution of partial differential equations (Spravochnik po chislennomu resheniya differentsial'nykh uravnenii v chastnykh proizvodnykh)*, 3rd edition, Gostekhizdat (1949).

2. Numerical solutions of boundary value problems associated with differential equations of elliptic type, *Usp. matem. nauk* 4 (1947).

PAPKOVICH P. F.

1. Structural mechanics of ships (Stroitel'nya mekhanika korablya), *Izd. sud. prom.* (1941).

PETROV G. I.

1. Application of the method of Galerkin to a problem involving the stationary flow of a viscous fluid, *Prikl. matem. i mekh.* **4**, 3 (1940).

PETROVSKY I. G.

1. A new proof of the existence of finite difference solutions of the Dirichlet problem, *Usp. Matem. Nauk*, VIII, (1940).

2. *Lectures on partial differential equations (Lektsii ob uravneniyakh s chastnymi proizvodnymi)*, Gostekhizdat (1950).

PICCONE M.

1. Nuovo metodo d'approssimazione per la soluzione del probleme di Dirichlet, *Rendiconti Acad. d. Lincei*, **5**, 31 (1922).

POL'SKY N. I.

1. On the convergence method of B. G. Galerkin, *Dopov. Akad. Nauk Ukr. SSR.* Bulletin of Physico Mathematical and Chemical Sciences, 6 (1949).

2. The convergence of some approximate methods of analysis, *Ukr. Matem. Zh.*, **7**, 1 (1955).

PRIVALOV I. I.

1. *Integral equations (Integral'nye uravneniya)*, ONTI DNTVU NKTP, (1935).

RAFAL'SON Z. KH.

1. A problem arising in the solution of the biharmonic equation, *Dokl. Akad. Nauk SSSR* **64**, 8 (1949).

2. A problem arising in the solution of the biharmonic equation, *Uch. zap. Len. gos. un-ta*, 144, *ser. matem. nauk* 23, 165–191 (1962).

RAYLEIGH, Lord

1. *The theory of sound*, Dover, New York (1945).

REPMAN YU. V.

1. A problem in the mathematical bases of Galerkin's method for solving problems on the stability of elastic systems, *Prikl. matem. i mekh.* **4**, 2 (1940).

RIESZ F. and B. SZÉKESFALVI–NAGY

1. *Leçons d'analyse fonctionnelle*, Budapest (1952). English translation published as *Functional Analysis*, Ungar, New York (1955).

RITZ W.

1. Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik, *J. reine angew. Mathematik*, **135**, 1 (1908) and also W. Ritz, *Gesammelte Werke*, Paris (1911).

2. Theorie der Transwersalschwingungen einer quadratischen Platte mit freien Rändern, *Gesammelte Werke*, Paris (1911).

SHERMAN D. I.

1. On solving plane static problems in the theory of elasticity under given external forces, *Dokl. Akad. Nauk SSSR*, **28**, 1 (1940).

SHEVCHENKO K. N.

1. Application of variational methods to the solution of problems in the theory of elasticity, *Prikl. matem. i mekh.*, **2**, 2 (1938).

SHURA-BURA M. R.

1. *The use of electrical networks in the solution of the Dirichlet problem by the finite difference equation which approximates for Laplace's equation (O reshenii konechno-raznostnogo uravneniya, approksimiruyushchego zadachu Dirikhle dlya uravneniya Laplasa, na elektricheskikh setkakh). Computational mathematics and computational practice (Vychislitel'naya matem. i vychislitel'naya tekhnika)* 1, Izd. Akad. Nauk. Moscow (1953).

SLOBODYANSKY M. G.

1. A procedure for the approximate integration of equations with partial derivatives and its application to problems in the theory of elasticity, *Prikl. matem. i mekh.*, **3**, 1 (1939).

2. Questions of space for prismatic body problems in the theory of elasticity, *Uch. zap. Mosk. gos. un-ta*, 39, Mekhanika (1940).

3. The estimation of error in quantities determined by solving linear problems by variational methods, *Dokl. Akad. Nauk SSSR*, **86**, 2 (1952).

4. The estimation of the error of the approximate solution of linear problems by variational methods and applications to two sided approximations for static problems in the theory of elasticity, *Prikl. matem. i mekh.*, **16**, 4 (1952).

5. The estimation of error of approximate solutions of linear problems, *Prikl. matem. i mekh.*, **17**, 2 (1953).

6. The approximate solution of linear problems by variational methods, *Prikl. matem. i mekh.*, **17**, No. 5 (1953).

7. Approximate solutions of some boundary value problems for elliptic differential equations and the estimation of error, *Dokl. Akad. Nauk SSSR*, **89**, 2 (1953).

8. The transformation of minimum functional problems into problems of mechanics, *Dokl. Akad. Nauk SSSR*, **91**, 4 (1953).

9. The construction of approximate solutions of linear problems, *Prikl. matem. i mekh.*, **19**, 5 (1955).

SMIRNOV V. I.

1. *A Course of Higher Matematics (Kurs vysshei matematiki)* Vol. III, part II, Gostekhizdat (1957). English translation published by Pergamon Press, Oxford (1964).

2. *A Course of Higher Mathematics (Kurs vysshei matematiki)* Vol. IV, 5th edition, Gostekhizdat (1958). English translation published by Pergamon Press, Oxford (1964).

3. *A Course of Higher Mathematics (Kurs vysshei matematiki)* Vol. V, Gostekhizdat (1960). English translation published by Pergamon Press, Oxford (1964).

SOBOLEV S. L.

1. *Partial Differential Equations of Mathematical Physics (Uravneniya matematicheskoi fiziki)*, 1st edition (1947), 2nd edition (1950), 3rd edition (1954), Gostekhizdat. English translation published by Pergamon Press, Oxford (1964).

2. Some applications of functional analysis to mathematical physics (Nekotorye prilozheniya funktsional'nogo analiza k matematicheskoi fiziki), *Izd. Len. gos. un-ta*, (1950).

SRETENSKY L. N.

1. *The theory of wave motion in liquids (Teoriya volnovykh dvizhenii zhidkosti)*, ONTI DNTVU NKTP, (1934).

SVIRSKY I. V.

1. The estimation of the accuracy of approximate methods in determining the frequency of oscillations, *Izv. Kazansk filiala Akad. Nauk SSSR, ser. fiz-matem. i tekh. nauk*, 3, 59–86 (1953).

TIKHONOV A. N. and A. A. SAMARSKY

1. *Equations of Mathematical Physics (Uravneniya matematicheskoi fiziki)*, Gostekhizdat (1951). English translation of 2nd edition (1953) published by Pergamon Press, Oxford (1963).

TIMOSHENKO S. P.

1. Elasticity Theory (Teoriya uprugosti) ONTI (1937).
2. Plates and Shells (Plastinki i obolochki) Gostekhizdat (1949).

TREFFTZ E.

1. Ein Gegenstück zum Ritz'schen Verfahren, Proceedings of the Second International Congress for Applied Mechanics, Zurich, 131–138 (1926).

2. Matematische Elastizitätstheorie, *Handbuch der Physik* VI, Springer, Berlin (1928).

TRICOMI F.

1. *Partial differential equations of mixed type*, Brown University, Rhode Island (1948). Translation from Italian.

VARVAK P. M.

1. The development and application of the method of networks to calculations on plates, Part 1 (1949), Part II (1952). *Izd. Akad. Nauk Ukr. SSR*, Kiev.

VISHIK M. I.

1. The method of orthogonal projections for self adjoint equations, *Dokl. Akad. Nauk SSSR*, **56**, 2 (1947).

2. The method of orthogonal projections for general self-adjoint equations, *Dokl. Akad. Nauk SSSR*, **58**, 6 (1947).

3. The method of orthogonal and direct projections in the theory of elliptic differential equations, *Matem. sborn.* **25** (67); 2 (1949).

4. The first boundary value problem for elliptic equations which are degenerate on the domain boundary, *Dokl. Akad. Nauk SSSR*, **93**, 1 (1953).

5. Boundary value problems for elliptic equations which are degenerate on the domain boundary, *Dokl. Akad. Nauk SSSR*, **93**, 2 (1953).

6. Boundary value problems for elliptic equations which are degenerate on the boundary, *Matem. sborn.*, **35** (77); 3 (1954).

VOROVICH I. I.

1. The existence of solutions in the non-linear theory of shells, *Izv. Akad. Nauk SSSR, ser matem.*, **19**, 4 (1955).

VVEDENSKAYA N. D.

1. A boundary value problem for equations of elliptic type which are degenerate on the domain boundary, *Dokl. Akad. Nauk SSSR*, **91**, 4 (1953).

WALSH J.

1. Über die Entwicklung einer analytischen Function nach Polynomen, *Math. Ann.* **96**, 430–436 (1926/7).

WEIERSTRASS K.

1. Über das sogenannte Dirichlet'sche Prinzip, *Matematische Werke v. K. Weierstrass*, **2** (1895).

WEYL H.

1. Das asymptotische Verteilungsgesetz der Eigenschwingungen eines beliebig gestalteten elastischen Körpers, *Rendiconti del Circolo Matematico di Palermo*, **39** (1915).

2. The method of orthogonal projections in potential theory, *Duke mathematical journal*, **7** (1940).

ZAREMBA S.

1. Sur le principe de minimum, *Bull. Internat. de l'Acad. des sciences de Cracovie, Classe des sciences mathématiques et naturelles, No.* 7, 197–264 (Juillet 1909).

2. Sur un problème toujours possible comprenant, à titre de cas particuliers le problème de Dirichlet et celui de Neumann, *J. Math. pures appl.* **6**, 127–163 (1927).

3. L'équation biharmonique et une classe remarquable de fonctions fondamentales harmoniques, *Bull. intern. de l'Acad. des Sciences de Cracovie*, 3 (1907).

4. Sur l'intégration de l'équation biharmonique, *Bull. intern. de l'Acad. des Sciences de Cracovie*, 1 (1908).

ERRATA

In this index the references in roman numerals to preliminary page numbers are incorrect. To obtain the correct reference, the number printed has to be reduced by four in each case.

DIAZ 340
Differential equation *see* Ordinary differential equations
Dirichlet
  principle §§ 88–9; xx, 398–9, 400
  problem §§ 50, 80, 87; 120–1, 138, 161, 359, 373, 402, 545, 547
Distance between functions 43, 289 (*see also* Norm)
Domain 6–8
  closed 8
  open 8
  stellate 367
  trapezoidal § 97

EIDUS 269
Eigenfunctions 208
  of symmetric operator § 30
Eigenvalues 207
  bounds of § 62
  of elliptic operators § 37
  of ordinary differential equation § 35
  problem Ch. V, VIII; § 68; 475, 483; connexion with vibrations and stability § 29
  rank of 209
  spectrum of 226
  of symmetric operator § 30
Eigenvector 558
Elastic stress tensor 356–9
Elastic support § 67; 136
Elasticity theory §§ 26, 91–3; 97, 178–82, 198–200, 547
  operator 267–70
  two-sided estimates in § 54
Elements
  of a field 20
  limiting 323, 357
  null 323
Embedding theorem *see* Sobolev
EMDE 541–2 *see also* Jahnke
Energy
  completeness 237
  convergence Ch. II § 8
  finite §§ 16, 17; 110
  method Ch. III, 365, 495–7 applications of Ch. IV
  norm 52
  orthogonality 237
  product 55
  theorems §31
Equation
  degenerate 483
  elliptic §§ 80, 81; 162, 353–4, 367, 393, 399, 547
  hyperbolic 546–8
  parabolic 546–8
Error
  estimates of Ch. VII
  due to perturbation § 63
  in elasticity theory § 54
EULER 546
  equation 253
Extension
  of an operator 313
  of a set 320

FADDEYEVA 428, 550–1, 555, 563
Field of definition 20, 312
  width of 469
Field of values 20, 312
FIKHTENGOL'TS 244
Finite differences, method of Ch. XI
  Laplace equation § 96; networks '94; Poisson equation § 96; straight lines §§ 95, 98; trapezoidal domain § 97
Fourier
  series 44, 59–73, 311, 337, 397, 510, 516
  coefficients 63, 301–4, 417, 452, 515
Fredholm
  alternative 465, 471
  equation § 74; 450, 475, 501–3
  operator 22, 26–7, 460, 479, 532, 534
Friedrichs xxx, 269–70, 321, 547 (*see also* Courant)
  inequality 146, 165, 192, 203
  transformation xxxvi, 333, 380
FUCHS 469
Function
  vector 335, 341–52
  holomorphic § 86, 465, 531
Functional §§ 4, 45; 25
  bilinear 372
  bounds of § 61
  linear 309
Fundamental sequence 300

GALERKIN xxxiv (*see also* Bubnov)
Gaus's theorem 9, 147, 166, 341
GEL'FAND 176, 268–9
Generalized (concepts of)
  derivative 105, 490, 506
  functional § 45
  integrability § 41
  limit process § 43
  operator § 45
  orthogonal projections § 51
  orthogonality § 44
  solution 125
  Trefftz's method §§ 57, 59
  variational problem § 46
GERMAIN 188
  equation 191, 419
GERSHGORIN 548
GLAZMAN 277 (*see also* Akhiyezer)
Gram determinant 57–9, 88, 229, 294, 492
Green's
  formulae 10–12, 329–30
  function 97, 363–4, 386, 388, 425–6, 478, 504, 505
GROSBERG 35 (*see also* Levin)

HADAMARD xxii
Heat transfer 3
HILBERT xxii, 333, 394 (*see also* Courant)
  space Ch. VI
  complex 348, 397
  functional § 42

Identity operator 22
Integrability *see* Lebesgue
  Riemannian 280–1
Integral equation § 84, 386 (*see also* Fredholm)
  singular 509
Integrated kernel 386, 426
Integration by parts 9

JAHNKE 541–2

KANTOROVICH xix, xxxvi, 85, 95, 161, 547–8
  method 98, 549
KATO 384–5
KELDYSH xxxv, 168, 476, 480, 483
KHARRIK 400
KIRCHHOFF 75, 183
KLIOT-DASHINSKY 340
KRASNOSEL'SKY xxxv, 212
KRAVCHUK xxxvi, 501, 503
KRYLOV xxxvi, 85, 95, 503, 547–8 (*see also* Kantorovich)

LADYZHENSKAYA 548
Lagrange's
  equation *see* Germain
  method of undetermined multipliers 228, 231
Lamé equations 177, 183, 356
LANGENBAKH 565
Laplace equation §§ 22, 47, 96; 1, 119, 151–60, 339, 359, 545, 548, 550, 566
Laplacian operator §§ 58, 83; 21, 34, 54–6, 114, 138–151, 257, 259, 262, 264, 397, 402, 411–13
Lauricella's equation 532
Least squares,
  method of Ch. X xxxv, xxxvi
  applications §§ 84, 85
  convergence 493
  Dirichlet problem §§ 87–9
  foundations § 83
  mixed problem § 90
  Neumann problem § 87
  periodic problem of elasticity theory § 92
Lebesgue integral § 41; 19, 281–7
  of complex function 285
Legendre polynomials 62
LEIBENZON 151–3, 176, 416
LEVI 547 (*see also* Courant)
LEVIN 35
Limit
  process in Hilbert space § 43
  of a sequence 295
Limiting element 298–9
LIN HUNG-SUN 159
Linear
  independence § 9; 294
  problem xxiv
Linearity 309
LIPSHITZ condition 521, 529
LYUSTERNIK 547

MAKHOVER 168
MARKUS 547
Measure of a point set 278–9

MELAN 547 (*see also* Bleikh)
Metric 289
MIKELADZE 548
MIKHLIN 321
Minimax
  principle § 40, 273–4
  property of eigenvalues 386
Minimizing
  function 219
  sequence § 13; 123, construction of § 15
Minimum functional 319, 370, 496
  problem § 19
  theorem § 11; 75, 179
Minimum
  potential energy, principle of § 26; 78, 178, 190, 359
  surface integral, method of § 47
Mixed problem §§ 47, 81, 90; 3, 140, 143, 161, 178–82, 356
MOREN 340
MUSKELISHVILI 521, 526

NATHANSON 101, 278, 305, 307
Networks
  electric 548
  method of xxxvi, § 94
Neumann problem §§ 53, 81, 87; 3, 141–3, 149–61, 340, 374, 393, 402
Newton's method 161
Norm 15, 309
  of operator 313
NOVOZHILOV 212
Null operator 22
Numerical examples Ch. VIII

OLEINIK 168
Operator §§ 4, 45, 29, 312
  biharmonic *see* Biharmonic
  bounded 99, 313
  bounded-below 218
  closeness of 392
  completely continous §§ 76–7, 470–5
  continuous 313
  degenerate 459, 461, 470–75
  differential 22
  elasticity theory 267–70
  elliptic 162, 256–9; eigenvalues of § 37
  extension of 313
  field of definition of 312
  field of values of 312
  finite *see* degenerate
  identity 314
  integral 22 (*see also* Fredholm)
  inverse, 29, 315, 461
  Laplace *see* Laplacian
  linear 23–4, 313
  null 314
  permutation of 29
  polyharmonic 399
  positive-bounded-below § 6, 36–9, 318
  positive definite § 6, 318, 350–2
  regularizing 500
  representation of as T^*T 350–2
  similar 401
  size of 274
  symmetric §§ 5, 30
  Tricomi 163, 169
Ordinary differential equation §§ 75, 79, 98; 98
  boundary value problem for § 20, 477
  eigenvalues of §§ 35, 68
  of even order 248–50
Orthogonal
  complement 338, 354
  component 368
  projections, method of xxxvi, §§ 49, 50, 51, 336, 342, 353–4, 359, 374, 384, 408–11
  series §§ 10, 12; 65–73, 302
  subspace 367
  sum 338
Orthogonality §§ 10, 44
  energy 237
  of spaces § 49; 337
  with weight function 70
Orthogonalization 60, 346–8
Orthonormal system 215
Orthonormalization 60, 71
  energy 67–8

PANOV xxxvi, 547–8
PAPKOVICH 200
Parseval equation §22; 66, 271, 304
Peturbation of an equation § 63
PETROV xxxv, 476
PETROVSKY 1, 547

PICCONE xxxvi, 510
Poincaré's inequality § 22; 149–50, 167
Poisson equation §§ 53, 58, 85, 96, 97; 4, 370, 389–9, 400, 402–8, 548, 562
POL'SKY xxxv, 476–7
Polyharmonic operator 399
Potential
  energy 186
  theory § 90
Product *see* Scalar *and* Energy
Projections § 49 (*see also* Orthogonal)

RAFAL'SON 365, 367, 369, 416
Rayleigh's method xxv
Regular values 465
REPMAN xxxv, 450
Resolvant 453–5, 477, 355
RIEMANN xx
  integrability 280–1
  integral 286
RIESZ 277
Riesz–Fischer theorem 101–2, 105, 286–7
RITZ xxii
  method §§ 4, 32, xxii 33, 85, 112–3, 137–8, 192, 204, 230, 328, 330, 332, 334, 344, 346, 348–50, 359, 385, 400–08, 410–19, 421–40, 445–8, 476–7, 497, 499, 504, 549
Rouche's theorem 469

SAMARSKY 35 (*see also* Tikhonov)
Scalar product
  of functions § 3; 13, 288
  continuity of 46
Sequence
  minimizing xxxi, §§ 13, 15, 123
  fundamental 300
Series, orthogonal §§ 10, 12; 65–73, 204, 302
Set
  close 72, 278
  compact 222
  complete 65
  dense 299
  enumerable 278, 307
  linear 23, 287, 312
  measure of a 278–9
  open 278
  orthogonal and orthonormal 60
  sum of 279
SCHIFFER 168 (*see also* Bergman)
Schwartz's
  formula 519
  kernel 525
SHABAT 469 (*see also* Fuchs)
SHERMAN 531
SHEVCHENKO 97
SHURA-BURA 548
Singularity, non-essential 254–9, 268–9
SLOBODYANSKY 334, 384–5, 550, 553, 566
  method § 60
SMIRNOV xvi, xvii, 277–8, 305, 425, 469, 495, 497
SOBOLEV xix, xxx, 105, 149, 193, 374
  embedding theorems 149, 193, 198, 205, 265, 374
SOMILYAN tensor 267
Space
  Hilbert Ch. VI
  H_η 293
  $\mathfrak{H}$ 341–52
  $L_2(\Omega)$ 291, 335, 367, 460, 481, 506
  $\mathbf{L}_2(\Omega)$ 293
  $L_2(\Gamma)(\Omega)$ 506
  $L_2(p, S)$ 513, 516
  separable 305–6
Spectrum, discrete, 226, 242–50
SRETENSKY 160
Stability §§ 29, 36, 38, 72; 211–12
Steepest descent, method of 99
Straight lines, method of §§ 95–8
Subspaces § 49
  complementary 343
Summability 285
  quadratic 287
SVIRSKY 386, 389, 391
Symmetry of an operator 317
SZÉKESFALVI-NAGY 277 (*see also* Riesz)

Tangential stress 404–13
TIKHONOV 35

TIMOSHENKO xxxii, xxxiii, 35, 151, 200, 404
Torque 404–13
Torsion
  of a rod §§ 23, 65; 81–3
  problem 562
Trace 386
TREFFTZ xxxvi, 140, 177, 267
  method §§ 55, 57–9, 333, 355, 365; application 411–13
Triangle inequality 17, 18. 102, 132, 289–90, 296
TRICOMI 168
  operator 163, 169
Trivial solution 207

Values (*see also* Eigenvalues)
  characteristic 465
  regular 465
Variable coefficients § 24
Variational problem
  existence of solution § 19; 21
  solutions of § 46
VARVAK 547
Vibrations
  of a beam § 69
  characteristic § 29
  cigen – *see* Characteristic
  of an elastic body § 39
  of a hollow truncated conc 429
  of a membrane 210–11
  of a plate § 71; 275–6
  radial, of a cylinder § 70
  of a solid cone 433–5
  of a string
VISHIK xxxvi 168, 174, 340, 354
VVEDENSKAYA 168

WALSH 505
  theorem 505, 510, 517, 529
WEIERSTRASS xx, 244
  theorem 306, 459
Weight function 53, 513, 516
WEYL 269, 340, 367

ZAREMBA xxxvi, 339–40, 365

OTHER TITLES IN THE SERIES IN PURE AND APPLIED MATHEMATICS

Vol. 1 WALLACE—*Introduction to Algebraic Topology*
Vol. 2 PEDOE—*Circles*
Vol. 3 SPAIN—*Analytical Conics*
Vol. 4 MIKHLIN—*Integral Equations*
Vol. 5 EGGLESTON—*Problems in Euclidean Space: Application of Convexity*
Vol. 6 WALLACE—*Homology Theory on Algebraic Varieties*
Vol. 7 NOBLE—*Methods Based on the Wiener-Hopf Technique for the Solution of Partial Differential Equations.*
Vol. 8 MIKUSINSKI—*Operational Calculus*
Vol. 9 HEINE—*Group Theory in Quantum Mechanics*
Vol. 10 BLAND—*The Theory of Linear Viscoelasticity*
Vol. 11 KURTH—*Axiomatics of Statistical Mechanics*
Vol. 12 FUCHS—*Abelian Groups*
Vol. 13 KURATOWSKI—*Introduction to Set Theory and Topology*
Vol. 14 SPAIN—*Analytical Quadrics*
Vol. 15 HARTMAN and MIKUSINSKI—*Theory of Lebesgue Measure and Integration*
Vol. 16 KULCZYCKI—*Non-Euclidean Geometry*
Vol. 17 KURATOWSKI—*Introduction to Calculus*
Vol. 18 GERONIMUS—*Polynomials Orthogonal on a Circle and Interval*
Vol. 19 ELSGOLC—*Calculus of Variations*
Vol. 20 ALEXITS—*Convergence Problems of Orthogonal Series*
Vol. 21 FUCHS and LEVIN—*Functions of a Complex Variable, Volume II*
Vol. 22 GOODSTEIN—*Fundamental Concepts of Mathematics*
Vol. 23 KEENE—*Abstract Sets and Finite Ordinals*
Vol. 24 DITKIN and PRUDNIKOV—*Operational Calculus in Two Variables and its Applications*
Vol. 25 VEKUA—*Generalized Analytic Functions*
Vol. 26 AMIR-MOÉZ and FASS—*Elements of Linear Spaces*
Vol. 27 GRADSHTEIN—*Direct and Converse Theorems*
Vol. 28 FUCHS—*Partially Ordered Algebraic Systems*
Vol. 29 POSTNIKOV—*Foundations of Galois Theory*
Vol. 30 BERMANT—*A Course of Mathematical Analysis, Part II*
Vol. 31 LUKASIEWICZ—*Elements of Mathematical Logic*
Vol. 32 VULIKH—*Introduction to Functional Analysis for Scientists and Technologists*
Vol. 33 PEDOE—*Introduction to Projective Geometry*

Vol. 34 TIMAN—*Theory of Approximation of Functions of a Real Variable*
Vol. 35 CSÁSZÁR—*Foundations of General Topology*
Vol. 36 BRONSHTEIN and SEMENDYAYEV—*A Guide-Book to Mathematics for Technologists and Engineers*
Vol. 37 MOSTOWSKI and STARK—*Introduction to Higher Algebra*
Vol. 38 GODDARD—*Mathematical Techniques of Operational Research*
Vol. 39 TIKHONOV and SAMARSKII—*Equations of Mathematical Physics*
Vol. 40 McLEOD—*Introduction to Fluid Dynamics*
Vol. 41 MOISIL—*The Algebraic Theory of Switching Structures*
Vol. 42 OTTO—*Nomography*
Vol. 43 RANKIN—*An Introduction to Mathematical Analysis*
Vol. 44 BERMANT—*A Course of Mathematical Analysis, Part I*
Vol. 45 KRASNOSEL'SKII—*Topological Methods in the Theory of Nonlinear Integral Equations*
Vol. 46 KANTOROVICH and AKILOV—*Functional Analysis in Normed Spaces*
Vol. 47 JONES—*The Theory of Electromagnetism*
Vol. 48 FEJES TÓTH—*Regular Figures*
Vol. 49 YANO—*Differential Geometry on Complex and Almost Complex Spaces*
Vol. 50 MIKHLIN—*Variational Methods in Mathematical Physics*
Vol. 51 FUCHS and SHABAT—*Functions of a Complex Variable, Vol. I*
Vol. 52 BUDAK, SAMARSKII and TIKHONOV—*A Collection of Problems on Mathematical Physics*
Vol. 53 GILES—*Mathematical Foundations of Thermodynamics*
Vol. 54 SAUL'YEV—*Integration of Equations of Parabolic Type by the Method of Nets*
Vol. 55 PONTRYAGIN *et al.*—*The Mathematical Theory of Optimal Processes*
Vol. 56 SOBOLEV—*Partial Differential Equations of Mathematical Physics*
Vol. 57 SMIRNOV—*A Course of Higher Mathematics, Vol. I*
Vol. 58 SMIRNOV—*A Course of Higher Mathematics, Vol. II*
Vol. 59 SMIRNOV—*A Course of Higher Mathematics, Vol. III, Part 1*
Vol. 60 SMIRNOV—*A Course of Higher Mathematics, Vol. III, Part 2*
Vol. 61 SMIRNOV—*A Course of Higher Mathematics, Vol. IV*
Vol. 62 SMIRNOV—*A Course of Higher Mathematics, Vol. V*
Vol. 63 NAIMARK—*Linear Representations of the Lorentz Group*

MADE IN GREAT BRITAIN